FIB

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Disseny de fàrmacs antiinflamatoris

Inhibidors de la COX-2

Juan José Vázquez Giménez

Titulació:	Grau en Enginyeria en Informàtica (Computació)
Director:	Luis Antonio Belanche Muñoz
Codirector:	Josep Maria Campanera Alsina
Data d'entrega:	June 21, 2016

Continguts

1 Resum

1.1 Resum

En aquest projecte s'ha desenvolupat un sistema capaç de fer recomanacions de millora per a dissenys de fàrmacs antiinflamatoris, concretament d'inhibidors de la COX-2. Les recomanacions es processen aprofitant la informació extreta d'un conjunt de dades que defineixen l'activitat de fàrmacs ja existents. Aquestes dades han estat obtingudes a partir de simulacions de la interacció de cada fàrmac amb la proteïna que intervé al procés.

Basant-se en un estat de l'art inicial s'ha descrit el procés de desenvolupament d'aquest sistema software. Explicant de forma detallada les decisions, tant tècniques com procedimentals, que s'han pres i el raonament que ens ha induït a cadascuna d'elles. Per concloure, s'han analitzat els resultats obtinguts i s'ha reflexionat sobre la qualitat de la solució proposada. Entre altres coses, això ens ha permès identificar possibles treballs futurs amb l'objectiu d'ampliar i millorar el sistema.

1.2 Abstract

In this project, it has been designed a system capable of making recommendations for the improvement of designs of anti-inflammatory drugs, concretely the COX-2 inhibitors. The recommendations are processed exploiting the information extracted from a data set that defines the activity

of existing drugs. This data has been obtained from simulations of the interaction of each drug with the protein involved in the process.

Based on an initial state of the art it described the development process of this software system. Explaining in detail the decisions taken, both technical and procedural, and the reasoning which led us to each. To conclude, we have analyzed the results and has reflected on the quality of the proposed solution. Among other things, this has allowed us to identify possible future work in order to expand and improve the system.

1.3 Resumen

En este proyecto se ha desarrollado un sistema capaz de realizar recomendaciones de mejora para diseños de fármacos antiinflamatorios, concretamente de inhibidores de la COX-2. Las recomendaciones se procesan aprovechando la información extraida de un conjunto de datos que definen la actividad de medicamentos ya existentes. Estos datos han sido obtenidos a partir de simulaciones de la interacción de cada fármaco con la proteína que interviene en el proceso.

Basándose en un estado del arte inicial se ha descrito el proceso de desarrollo de este sistema software. Explicando de forma detallada las decisiones, tanto técnicas como procedimentales, que se han tomado y el razonamiento que nos ha inducido a cada una de ellas. Para conluir, se han analizado los resultados obtenidos y se ha reflexionado sobre la calidad de la solución propuesta. Entre otras cosas, esto nos ha permitido identificar posibles trabajos futuros con el objetivo de ampliar y mejorar el sistema.

Part I

Primera part

2 Introducció

Aquest document conté la memòria del Treball de Fi de Grau en Enginyeria Informàtica amb títol: "Disseny de fàrmacs antiinflamatoris. Inhibidors de la COX-2". El grau s'imparteix a la Facultat d'Informàtica de Barcelona (FIB [2]) de la Universitat Politècnica de Catalunya (UPC [3]).

El contingut d'aquest capítol introductori és el següent:

1. **Problema:** Es defineix el problema que ha motivat la realització d'aquest projecte. Enumerant les raons per les quals és un problema i els avantatges de resoldre'l.
2. **Objectius:** Es plantegen quins objectius pretén assolir el treball. Raonant per què són els adequats a la qüestió que es vol resoldre i a les característiques del projecte.
3. **Abast:** Es perfila l'abast del projecte, així com quins possibles obstacles poden interferir a la realització de la tasca.
4. **Parts interessades:** S'enumeren les persones interessades que intervenen al projecte. Explicant com i per quines raons ho fan.

2.1 Problema

En aquesta secció es defineix el problema que ha motivat la realització d'aquest treball. Per facilitar la seva entesa, s'ha organitzat de la

següent manera:

1. **Context:** Es dóna una breu base teòrica d'alguns conceptes necessaris per a la correcta comprensió de la secció i de tot el document.
2. **Procés de disseny:** Es defineix el procés de disseny d'un fàrmac, explicant en quines fases es divideix i en què consisteix cadascuna.
3. **Concreció del problema:** Es concreta quin és el problema que aquest projecte pretén resoldre.

2.1.1 Context

La ciclooxigenasa-2, també coneguda com a PTGS2 i COX-2, és una proteïna del cos humà. La seva principal funció és intervenir als processos inflamatoris provocats per diverses causes:

- *Agents biològics:* bacteris, virus, fongs, paràsits.
- *Agents físics:* fred, calor, raigs UV, radiacions.
- *Agents químics:* Verins, toxines.
- *Altres:* alteracions vasculars o immunitàries, traumatismes i cossos estranys, i l'estrès.

Fora del focus inflamatori, la COX-2 també presenta altres efectes al cos humà [4] que s'enumeren a continuació:

- *Efectes sobre la funció renal*
- *Efectes gastrointestinals*
- *Efectes sobre el sistema cardiovascular*
- *Efectes sobre el sistema nerviós central*

- *Efectes sobre l'os*
- *Efectes en el càncer colorectal*
- *Efectes sobre l'úter i l'ovari*

Tornant a l'àmbit inflamatori, és important saber com la proteïna intervé en aquest tipus de processos. El fenomen que desencadena una inflamació és la intervenció de l'àcid araquidònic al centre actiu de la COX-2 [5]. Una forma d'inhibir aquest efecte, i la que nosaltres utilitzem, és utilitzant un lligand capaç d'evitar-ho. Segons inclou la Fundación Pública Andaluza para la Investigación Biosanitaria en Andalucía Oriental (FIBAO [6]) al seu glossari de medicina molecular [7], un lligand és una molècula capaç de ser reconeguda per una altra provocant una resposta biològica. D'aquesta manera, s'utilitza un lligand en forma de fàrmac capaç de ser reconegut per la COX-2 i evitar la interacció d'aquesta amb l'àcid araquidònic.

En el nostre cas, la funció que ha de complir el lligand és la d'ocupar el centre actiu de la proteïna. D'aquesta forma, impossibilita la intervenció de l'àcid araquidònic amb aquest centre. Perquè el lligand acompleixi la seva funció, és imprescindible que es posicioni al lloc adequat. A la Figura 2.1 podem veure l'exemple d'un lligand localitzat correctament sobre una estructura molecular, la qual podria ser una proteïna, en una simulació.

Per aconseguir que el lligand es posicioni correctament i assoleixi el seu objectiu s'ha de maximitzar la seva constant d'equilibri K_{eq}, que és el mateix que maximitzar la constant d'associació K_a. I per tant, minimitzar la seva constant de dissociació K_d, ja que es compleix: $K_{eq} = K_a = 1/K_d$. Mitjançant la funció de Gibbs: $\Delta G = -RTlnK_a$ es pot realitzar la conversió entre energia experimental i constant d'associació. En aquesta fórmula, ΔG representa l'energia experimental que uneix el lligand i la proteïna en kcal/mol, R i T són constants i K_a és la constant d'associació. A la Figura 2.2, poden veure una taula amb alguns exemples de conversió entre constant d'associació i energia d'associació.

Figura 2.1: Simulació d'un lligand localitzat correctament sobre una estructura molecular.

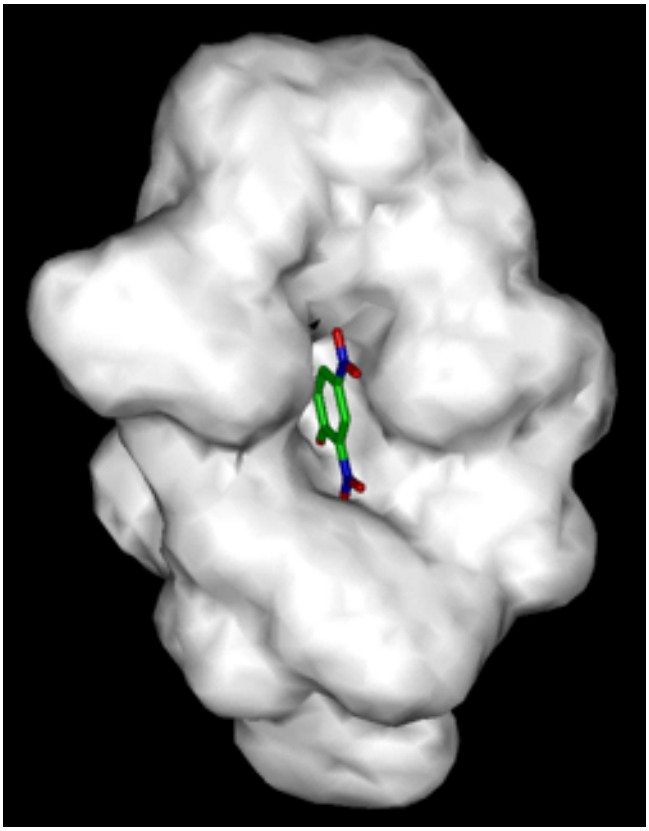


Figura 2.2: Taula de conversions entre constant d'associació i energia d'associació.

K_a	ΔG^o (kcal/mol)
10^{12}	-17.0
10^9	-12.8
10^6	-8.5
10^3	-4.3
10	-1.4
1	0

2.1.2 Procés de desenvolupament de fàrmacs

Una vegada contextualitzat el problema, estudiem el procés de desenvolupament d'un fàrmac [8], les seves fases i en què consisteix cadascuna.

El procés de desenvolupament de fàrmacs, defineix quines fases s'han de seguir des de la detecció d'una malaltia fins al llançament d'un fàrmac que la combat. A la figura 2.3 podem veure un esquema d'aquestes fases. A continuació, s'expliquen aquestes fases:

1. **Identificació de dianes:** Per començar, és necessari conèixer els mecanismes, rutes moleculars i proteïnes implicades a la malaltia. S'anomenen dianes als components del cos humà que en ser manipulats o modificats poden tenir un efecte positiu al curs de la malaltia.

2. **Obtenció de candidats inicials:** A continuació, se selecciona de forma definitiva la diana que mostra evidències del fet que és rellevant per la malaltia tractada. S'identifica quina és la millor forma de manipular-la per obtenir els resultats esperats. En molts casos, els nous

medicaments són compostos capaços de modificar el comportament de la diana interactuant amb ella; però hi ha altres casos on s'utilitza la diana com a medicament, aquest seria el cas de la utilització de la insulina en pacients diabètics. Si és el cas que el medicament serà un compost, és en aquesta fase on es defineixen els primers candidats.

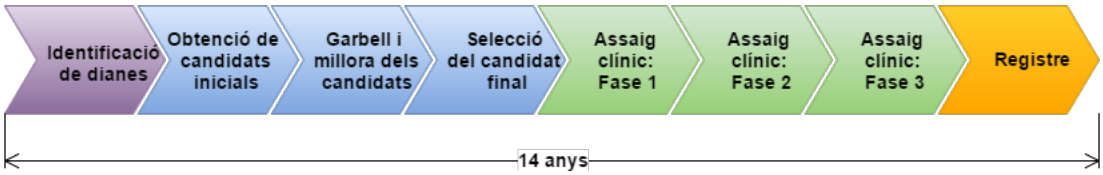
3. **Garbell i millora dels candidats:** En aquesta tercera fase s'estudien amb més detall els candidats inicials. Això permet eliminar els que presenten un efecte menys favorable o que mostren alguna raó per la qual la seva utilització no és recomanable. A banda d'això, s'intenta millorar la composició dels compostos seleccionats per potenciar el seu efecte.

4. **Selecció del candidat final:** Arribats a aquest punt, s'ha de proporcionar l'evidència que el nou fàrmac resultarà segur quan es provi en humans. Per proporcionar aquesta evidència, es realitzen un conjunt de proves in vitro i in vivo, aquestes últimes realitzades amb animals. Això permet obtenir informació més concreta sobre l'eficàcia dels compostos, donant la possibilitat de continuar millorant-los o seleccionar un com a candidat final per realitzar el seu estudi amb persones.

5. **Assaig clínic. Fase 1:** Una vegada seleccionat el candidat final, s'inicien els estudis amb humans. L'objectiu d'aquesta primera fase d'assaig clínic és confirmar que el fàrmac no presenta efectes perjudicials per a les persones. Per això, se subministra una petita dosi del compost a un grup reduït de voluntaris sans (entre 20 i 100). Si els resultats són favorables, es pot augmentar la dosi gradualment per determinar la seguretat del fàrmac en dosis més grans. Durant la totalitat d'aquesta fase, els voluntaris estan monitorats per assegurar la seva salut.

6. **Assaig clínic. Fase 2:** En aquesta segona fase d'assaig clínic, s'ha de comprovar l'eficàcia del fàrmac en un grup més extens (entre 50 i 250) de pacients que pateixen la malaltia que es vol tractar. En aquesta fase també s'administren diferents dosis per determinar quina és la més beneficiosa. Tot i que no és el principal objectiu, també s'estudia la seguretat del compost i la capacitat dels pacients de tolerar-lo. Per concloure aquesta fase, s'ha de definir una dosi que balancegi de la

Figura 2.3: Fases del procés de desenvolupament d'un fàrmac.



forma més favorable possible els efectes positius sobre la malaltia i la màxima reducció dels efectes negatius.

7. **Assaig clínic. Fase 3:** Per acabar amb l'assaig clínic, es realitza una última prova amb pacients. Aquest últim assaig té la finalitat de provar amb un grau de certesa elevat que el fàrmac és eficaç i segur. Per aconseguir aquest grau de certesa major, s'augmenta el nombre de pacients de forma significativa (aproximadament 500). Normalment, en els estudis d'aquesta fase es compara el nou fàrmac amb el qual actualment s'utilitza per competir la malaltia. Si no existeix cap fàrmac per aquesta, es compara amb un placebo per demostrar que realment té un efecte positiu en els pacients. En qualsevol dels dos casos anteriors, el que es pretén demostrar és que el nou fàrmac millora el tractament actualment més comú (que pot ser inexistent). Normalment, és necessari realitzar un assaig d'aquestes característiques almenys dues vegades per donar la fase com ha superada.

8. **Registre:** Per últim, cal el vistiplau de les autoritats sanitàries abans de poder comercialitzar el compost. Per això, es revisa que el procés s'ha realitzat d'una forma correcta i independent. D'aquesta manera, s'assegura que només es comercialitzen els fàrmacs que realment són eficaços i segurs. L'ultima decisió, només la poden prendre les autoritats sanitàries.

Segons el Jones Research Group [9] de la Universitat Estatal de Washington (WSU [10]), el procés sencer de desenvolupament d'un fàrmac té un cost aproximat d'uns 739 milions de dòlars i una duració d'entre 14 i 15 anys. D'aquest cost, els 6 anys que representen les 3 primeres fases del

procés, signifiquen uns 230 milions de dòlars. Això és més d'un 30% del cost total.

Per altra banda, sabem que en aquestes primeres fases és en les que més possibilitats hi ha que el procés fracassi. Això és així perquè cada fase superada intenta descartar possibles raons per les quals el fàrmac pot demostrar ser inapropiat per a complir la seva finalitat.

2.1.3 Concreció del problema

Una vegada contextualitzat el problema i conegut com es desenvolupa un nou fàrmac, podem concretar una mica més quin és el problema real que hem detectat. Donades les característiques de cada fase del procés, ens adonem del fet que les 3 primeres són les que, a priori, més possible sembla que es puguin ajudar a resoldre amb alguna eina computacional d'anàlisi de dades. Ja que les altres presenten dependències a causa de les proves amb animals i persones i a la revisió de les agències sanitàries. Per tant, centrem els nostres esforços en aquestes primeres fases.

Com ja hem dit, aquestes primeres fases representen una part important dels costos temporals i econòmics i són les que més probabilitats de fracassar tenen. Això implica que un gran nombre de procediments acaben fallant, el que significa una quantitat important de recursos econòmics i temporals que acaben per no ser quasi productius.

2.2 Objectius

El principal objectiu del present Treball Final de Grau és ajudar, en la mesura del possible, a resoldre el problema que hem detectat a l'apartat anterior. Aquest és bàsicament el fet que les 3 primeres fases del procés de desenvolupament d'un fàrmac representen aproximadament un 30% dels recursos econòmics, un 40% dels temporals i són les que més probabilitats de fracassar tenen. Això implica que un gran nombre de procediments acaben fallant, el que representa una quantitat important de recursos econòmics i

temporals que acaben per no ser quasi productius. Nosaltres ens centrem en el desenvolupament d'un tipus concret de fàrmacs que són els fàrmacs antiinflamatoris inhibidors de la COX-2.

Amb la finalitat de reduir l'impacte d'aquest problema, hem detectat dues possibilitats:

1. **Reducció de costos:** La primera possible orientació és la reducció de costos, tant econòmics com temporals, del procés de desenvolupament dels inhibidors de la COX-2. Aquestes són les conseqüències típiques de l'automatització de tasques mitjançant eines de software. El desenvolupament d'una nova eina de software concentra la majoria del seu cost a la seva producció. També pot representar un cost el procés d'aprenentatge de l'usuari i el manteniment d'aquesta, però és molt menys significatiu que el primer. A banda d'això, la capacitat d'execució d'un humà sol ser molt més lenta que la d'un computador. El que es tradueix en una millora temporal important. Perquè obtenir aquests avantatges sigui possible, la tasca ha de complir dos requisits fonamentals. El primer és que es tracti d'una tasca on la seva lògica sigui traslladable a un producte software. L'altre requisit a complir, és que la tasca sigui repetible en el temps. Tot i que hi ha casos en el que la tasca és tan crítica que val la pena desenvolupar un software per resoldre-la un número petit de vegades, l'amortització econòmica i temporal del desenvolupament d'una eina computacional sol dependre en gran mesura de les vegades que la tasca hagi de ser repetida en el temps.

2. **Increment de l'efectivitat:** L'altra forma d'afrontar aquest problema és incrementant la fiabilitat d'aquestes fases. Si es disminueix el percentatge de fallada del procés, sense, o gairebé sense incrementar ni el cost econòmic ni el temporal, s'està reduint els casos de fracàs i per tant la quantitat de temps i diners que no retribueixen amb un resultat satisfactori. Segons quin sigui l'expectativa d'aquest enfocament, és notablement més complex. Si es pretén desenvolupar una eina capaç de prendre millors decisions que un especialista, als requisits que ja presentava la primera possibilitat se li suma el fet que partint del coneixement de l'especialista s'ha d'aconseguir millorar els resultats envers aquest mateix professional. L'altra variant, és dotar al científic d'informació

útil, de la qual manca ara o és altament complexa d'aconseguir, per permetre que les seves decisions incrementin el seu nivell d'encert.

Les dues possibilitats no són exclusives entre elles. A causa d'aquest fet, hem pres la decisió de seguir una combinació d'ambdues. Per una banda, es vol desenvolupar una eina que acceleri l'extracció d'informació. Aquesta tasca és necessària a l'hora d'identificar els primers candidats, realitzar garbells i seleccionar els millors compostos. Per altra banda, l'anàlisi realitzada per la computadora pretén ser més exhaustiu del que seria portat a terme per una persona. Aquesta anàlisi pretén esgotar les possibilitats i donar informació molt més completa. Amb aquesta finalitat, es defineixen els següents objectius generals:

- **Extreure informació:** Extreure informació útil a partir de l'anàlisi de les dades de les quals disposem.
- **Representar gràficament la informació:** Representar amb gràfics de diferents tipus la informació extreta.
- **Fer suggeriments per la millora d'inhibidors de la COX-2:** Fer suggeriments de quins residus és recomanable potenciar o debilitar per afavorir la unió total del fàrmac amb la proteïna.
- **Oferir una interacció adequada:** Desenvolupar algun tipus d'interfície simple i funcional per facilitar l'ús de l'eina als usuaris.

2.3 Abast

En aquest apartat es defineix l'abast del projecte i quins imprevistos poden modificar la planificació inicial.

2.3.1 Definició de l'abast

En primer lloc, definir l'abast. Per adaptar aquesta part de la documentació a la metodologia emprada, hem dividit la feina en blocs suficient-

ment diferenciats però sense definir els requisits concrets i definitius, ja que poden canviar durant el desenvolupament del treball. Aquests blocs, ordenats de forma cronològica, són els següents:

1. **Estat de l'art:** Aquesta ha de ser la primera feina a realitzar. Consisteix a investigar, descobrir i valorar quines tècniques s'utilitzen al mercat actualment. Això permet tenir una visió inicial de quines opcions tenim disponibles, la seva complexitat i els resultats que poden donar.

2. **Selecció d'eines de desenvolupament:** Abans d'iniciar el desenvolupament del projecte, és indispensable valorar les alternatives tecnològiques existents al mercat i decidir quines d'aquestes s'adeqüen més al projecte en qüestió. Aquesta fase pot requerir proves de concepte, si la tecnologia no es coneix anteriorment.

3. **Familiarització amb les dades i l'entorn:** Per començar amb el desenvolupament del projecte, és imprescindible conèixer de quines dades disposem i quin és el seu format. Així com familiaritzar-se amb l'entorn seleccionat per tal d'agilitzar i millorar el desenvolupament del producte final.

4. **Preprocés de les dades:** Una vegada conegut quin és el format de les dades, s'ha de reflexionar sobre si aquest és el format idoni o no, i en aquest segon cas, definir quin ho és. Això pot incloure des de canvis en el tipus de fitxer emprat fins a neteja de dades que no aporten informació però pot alentir el còmput.

5. **Extracció d'informació:** A continuació, hem d'extreure informació de les dades de les quals disposem. Aquesta tasca inclou una estreta col·laboració entre el desenvolupador i l'especialista, ja que és el qui sap quines dades són més significatives i més útils. L'extracció d'informació consisteix a, a partir d'una gran quantitat de dades difícilment interpretables, donar un conjunt reduït de dades concretes capaces de resumir i destacar la informació útil que aporten.

6. **Representació de la informació:** Per facilitar la interpretació de la informació estreta per part de l'usuari, es generen un conjunt de gràfiques que la mostren d'una forma visual, agradable i funcional. Per

això, és necessari seleccionar quin tipus de gràfic s'adapta millor a la informació requerida.

7. **Generació de recomanacions:** Amb l'objectiu d'assistir a l'usuari en el procés de desenvolupament dels inhibidors de la COX-2, implementem la funcionalitat de recomanació de millores. Aquesta consisteix en el fet que la computadora sigui capaç de, analitzant tota la informació extreta prèviament, recomanar a l'especialista quins residus són més favorables o desfavorables per a l'objectiu final: maximitzar la unió entre el lligand i la COX-2.

8. **Desenvolupament d'una aplicació web:** Per últim, volem donar a l'usuari la possibilitat d'interactuar amb el sistema d'una forma senzilla, agradable i funcional. Això es fa amb la finalitat que la utilització d'aquest sistema requereixi el procés mínim d'aprenentatge.

2.3.2 Possibles obstacles

Després de definir quins són els passos a seguir en el desenvolupament del projecte, és interessant repassar quins obstacles podem trobar i quina seria la forma adient de superar-los. Els possibles obstacles que hem detectat són els següents:

- **Canvi de requisits:** En un projecte dirigit amb metodologies àgils, és possible que els requisits inicials vagin variant al llarg del procés. Això és conseqüència de dues característiques d'aquestes metodologies. La primera és que no és obligatori que a l'inici del projecte tots els requisits estiguin completament definits, el que pot derivar en estimacions poc encertades. L'altre, és que aquest tipus de metodologies accepten el canvi, fins i tot al final del projecte. Aquesta característica ajuda al producte a adaptar-se al mercat de forma rapida i eficaç, però fa molt difícil que el resultat final encaixi totalment amb els objectius que aquesta documentació requereix definir. Per intentar evitar aquest obstacle, hem decidit desenvolupar una primera versió que contingui el mínim de funcionalitats possible. D'aquesta forma, obtindrem una primera versió molt primerenca i una vegada acabada, s'aniran afegint millores contínuament.

- **Temps insuficient:** Continuant a la línia del primer obstacle, els projectes àgils es poden tancar per funcionalitats disponibles al producte final o per temps de desenvolupament, però no pels dos aspectes alhora. En el nostre cas, la limitació temporal és clarament la data límit de lliurament del projecte. Però, com s'han definit els objectius del projecte de forma inicial en aquesta documentació, d'alguna manera també s'estableix quines funcionalitats ha de contenir. A banda de la solució que hem definit per al primer obstacle, també es planteja l'opció del fet que finalment només es desenvolupi el nucli de l'aplicació deixant com a tasca futura la interfície. Tot i que aquesta situació és poc probable.

- **Pèrdua de material:** Un altre possible risc és la pèrdua de material d'algun tipus (dades, codi i documentació). Aquest fenomen pot ser causat per la fallada d'algun equip o per una mala utilització d'aquest. La forma més comuna de resoldre aquestes situacions és mitjançant l'ús de còpies de diferents tipus al núvol. Per a les dades, s'utilitza un servei d'emmagatzematge en línia que ens permet mantenir una còpia de seguretat en cas de necessitar-la. A més, permet l'accés a les dades des de qualsevol equip sense un suport físic. En el cas del codi, s'utilitza un controlador de versions que, entre moltes altres coses, ens facilita els mateixos avantatges que el servei d'emmagatzematge digital. Per la documentació, l'estratègia és una mica diferent. Com que s'edita utilitzat un editor en línia, el lloc on es troba la versió més actual és al mateix editor. En aquest cas, es fan còpies locals amb freqüència per mantenir una còpia de seguretat actualitzada.

- **Indisposició d'algun membre de l'equip:** Aquest és l'obstacle més difícil d'anticipar. La situació pot variar molt depenent de qui sigui el membre de l'equip que no pugui participar en el desenvolupament del projecte. A continuació plantegem quin és l'efecte per cada un dels membres de l'equip:

  - *Director del projecte:* Donat que el director del projecte aporta un perfil expert en les tècniques utilitzades, la seva absència es veuria traduïda en una reducció de la qualitat del producte final. Tot i això, l'autor intentaria compensar aquesta mancança amb més dedicació a la part teòrica i hauria de reduir les seves hores de desenvolupament.

  - *Codirector del projecte:* En aquest cas, el codirector és la persona que coneix l'àmbit per al qual es desenvolupa l'eina. Sense la seva col·laboració seria molt difícil adaptar el sistema al seu objectiu final. El desenvolupament es basaria en el plantejament inicial, fet que impossibilitaria la capacitat d'adaptació i evolució continua cap al mercat final.
  - *Autor del projecte:* L'autor del projecte és el perfil més tècnic de l'equip i, per tant, qui executa les decisions preses per l'equip. Amb la seva absència seria impossible la realització del projecte, ja que aquest pretén provar la seva capacitat per portar a terme un treball científic. Depenent de la durada de la seva absència es podrien seleccionar diferents solucions, com: la reducció de l'abast del projecte, la redistribució temporal de la càrrega de treball o, en el pitjor dels casos, l'ajornament del lliurament.

- **Insuficiència de dades:** Per últim, ens podem trobar amb un problema d'insuficiència de dades. Aquest fenomen pot influir directament amb el resultat final del projecte. Els conjunts de dades utilitzats en el desenvolupament d'aquest software ve donat pel codirector d'aquest. Així doncs, si es detecta una mancança, es podria sol·licitar un major volum d'aquestes. Si això no fos possible, la qualitat del producte final no serà l'esperada. Tot i així, l'eina seria desenvolupada i una feina futura seria provar-lo per a un conjunt de dades adient.

3 Conceptes previs

En aquesta secció s'exposaran alguns conceptes necessaris per a la correcta comprensió del document. En ordre d'aparició, aquests són els conceptes que s'explicaran:

1. **Tipus d'energia:** Es parla sobre quins tipus d'energia es contemplen en aquest estudi i que vol dir cadascun.
2. **Mecànica molecular i mecànica quàntica [11]:** Es fa una petita introducció a aquests dos conceptes.
3. **Coeficients de correlació:** S'expliquen els coeficients de correlació que s'han utilitzat i les seves característiques.
4. **Variància explicada i no explicada:** Es comenten que són cadascuna i quina és la diferència entre totes dues.

3.1 Tipus d'energia

A la nostra col·lecció de dades, disposem de l'energia d'interacció entre el lligand i cadascun dels residus que conformen la COX-2. Aquesta energia està descomposta segons uns tipus d'energia. Aquest tipus, poden ser independents (que no depenen de cap dels altres) o combinacions d'independents. A continuació, s'expliquen els 9 ordenats de més independents a menys i amb l'abreviació entre parèntesis:

- **Energia interna (int):** L'energia interna és l'energia associada a les distàncies, angles i diedres dels enllaços covalents. En el nostre cas particular aquesta contribució és irrellevant, ja que durant els càlculs de l'energia d'unió es cancel·la.

- **Energia electrostàtica (ele):** Donades les dues càrregues puntuals q_1 i q_2 separades per una distància r. L'energia electrostàtica es calcula mitjançant la següent fórmula donada per la llei de Coulomb [12]:

$$U = K\frac{q_1 q_2}{r}$$

$$K = \frac{1}{4\pi\varepsilon_0} = 9 \cdot 10^9 Nm^2C^{-2}$$

$$\varepsilon_0 = 8.854 \cdot 10^{-12} C^2 N^{-1} m^{-2}$$

- **Energia de Van der Waals (vdw):** És l'energia d'estabilització molecular. Formen un enllaç químic no covalent en el qual participen dos tipus de forces o interaccions, les forces de dispersió i les forces de repulsió entre les capes electròniques de dos àtoms contigus.

- **Energia de solvatació polar (cal):** L'energia de solvatació polar és la part de l'energia de solvatació total que té caràcter polar.

- **Energia de solvatació apolar (sur):** L'energia de solvatació polar és la part de l'energia de solvatació total que té caràcter apolar.

- **Energia en gas (gas):** Es tracta de l'energia resultant de la suma de l'energia electrostàtica i l'energia de Van der Waals. O el que és el mateix:

$$gas = ele + vdw$$

- **Energia de solvatació (sol):** La solvatació és el procés d'associació de molècules d'un dissolvent amb molècules o ions d'un solut. En dissoldre's els ions en un solut, es dispersen i són envoltats per molècules de solvent. La força electrostàtica entre el nucli de l'ió i la molècula del solvent disminueix en gran mesura amb la distància entre la molècula de solvent i el nucli de l'ió. Així, l'ió més gran s'uneix fortament amb el solvent i a causa d'això, s'envolta d'un gran nombre de molècules de solvent. L'energia de solvatació és l'energia alliberada quan els ions de

la xarxa del sòlid s'associen amb les molècules del solvent. Compleix la relació:

$$sol = cal + sur$$

- **Energia polar (pol):** Els enllaços formats per àtoms diferents amb grans diferències d'electronegativitat, formen molècules polars. La molècula és elèctricament neutra al seu conjunt per tenir la mateixa quantitat de partícules positives i negatives, però no existeix simetria a la distribució de l'electricitat. Les molècules tals que els seus centres de càrregues positives no coincideixen amb les càrregues negatives, es denominen molècules polars. S'anomena enllaç polar si un parell d'electrons de la configuració electrònica externa no està igualment compartit pels dos àtoms. Està composta per l'energia electrostàtica i l'energia de solvatació polar:

$$pol = ele + cal$$

- **Energia apolar (npol):** Les molècules apolars són aquelles molècules que es produeixen per la unió entre àtoms que posseeixen la mateixa electronegativitat. Per aquesta raó, les forces amb les quals els àtoms conformen la molècula atrauen els electrons de l'enllaç, són iguals, produint-se així l'anul·lació d'aquestes forces. Altra opció és que s'originin molècules apolars per la unió entre àtoms amb diferents electronegativitats; si la molècula resultant té una geometria regular a causa de la inexistència de parts no enllaçants, els moments dipolars s'anul·laran a la suma vectorial pel que la molècula serà apolar amb un moment dipolar total nul. L'energia apolar és l'energia a la qual intervé alguna molècula apolar. Està formada per l'energia de Van der Waals i l'energia de solvatació apolar:

$$npol = vdw + sur$$

- **Energia lliure d'unió (total):** L'energia lliure d'unió és l'energia total que relaciona el residu amb el lligand. Es pot expressar com la suma de les energies anteriors:

$$total = ele + vdw + cal + sur = gas + sol = pol + npol$$

3.2 Mecànica molecular i mecànica quàntica

A la **mecànica molecular**, les equacions són utilitzades per seguir les lleis de la física clàssica i aplicar-les al nucli molecular considerant als electrons fora. Bàsicament, la molècula és tractada com una sèrie d'àtoms connectats per enllaços. Les equacions, derivades de la mecànica clàssica, són utilitzades per calcular diferents interaccions i energies (camps de forces) provocades per allargaments d'enllaços, angles corbats, energies de torsió i interaccions sense enllaços. Aquests càlculs requereixen un conjunt de dades o paràmetres que descriuen interaccions entre diferents conjunts d'àtoms. Les energies calculades per la mecànica molecular no representa una quantitat absoluta perquè és utilitzada quan es compara diferents conformacions de la mateixa molècula. Aquesta mecànica requereix un menor temps computacional que la mecànica quàntica. La mecànica molecular s'utilitza en les següents operacions o càlculs:

- Energia de minimització.
- Identificació de conformacions estables.
- Càlcul de l'energia per conformacions específiques.
- Generació de noves conformacions.
- Estudi del moviment molecular.

La **mecànica quàntica** utilitza la física quàntica per a calcular les propietats d'una molècula per les interaccions considerades entre els electrons i els nuclis de les molècules. Perquè els càlculs siguin possibles, s'han de fer diverses aproximacions. En primer lloc, el nucli és observat en condicions d'immobilitat. És raonable després del moviment dels electrons, és més fàcil comparar. Després es consideren els electrons movent-se al voltant del nucli fix, és possible descriure l'energia electrònica separada de l'energia nuclear. En segon lloc, s'assumeix que els electrons es mouen independentment dels altres i així la influència d'altres electrons i nucli es pren com a mitjana. Els mètodes de mecànica quàntica s'adeqüen per fer els següents càlculs:

- Energia de l'òrbita molecular i coeficients.
- Calor de formació per conformacions específiques.
- Càrregues parcials atòmiques calculades dels coeficients de l'òrbita molecular.
- Potencials electrostàtiques.
- Moment dipolar.
- Estat de transició geomètrica i energètica.
- Energia de dissociació d'enllaços.

3.3 Coeficients de correlació

A probabilitat i estadística, la correlació [13] indica la força i la direcció d'una relació lineal i la proporcionalitat entre dues variables estadístiques. Es considera que dues variables quantitatives estan correlacionades quan els valors d'una d'elles varien sistemàticament respecte als valors homònims de l'altre: si tenim dues variables (V_1 i V_2) existeix correlació si en augmentar els valors de V_1 ho fan també els de V_2 i a l'inrevés. Això no implica cap relació de causalitat.

En aquest projecte parlarem de dos coeficients de correlació diferents:

- **Coeficient de Pearson:** El coeficient de correlació de Pearson és una mesura de la relació *lineal* entre dues variables aleatòries quantitatives. Aquest coeficient és independent de l'escala de mesura de les variables. Es calcula amb la següent fórmula:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  On X i Y són dues variables aleatòries quantitatives, σ_{XY} és la covariància de X i Y i, σ_X i σ_Y, són la desviació típica de cada variable.

- **Coeficient de Spearman:** El coeficient de correlació de Spearman és una mesura de la relació *monòtona* (sigui lineal o no) entre dues variables aleatòries quantitatives. Aquest coeficient és independent de l'escala de mesura de les variables. Es calcula amb la següent fórmula:

$$\rho_{X,Y} = 1 - \frac{6 \sum_{i=1}^{N} rg(X_i) - rg(Y_i)}{N(N^2 - 1)}$$

  On X i Y són dues variables aleatòries quantitatives ordenades creixentment, N és la mida de X i Y i, $rg(X_i)$ és la funció que denota la posició al conjunt original (sense ordenar) del valor de X_i.

En ambdós casos, ρ és un valor entre -1 i 1. Els extrems denoten una correlació perfecta, tant negativa com positiva respectivament, i 0 expressa no correlació, el que no vol dir que siguin independents.

3.4 Variància explicada i no explicada

La **variància** [14] d'una població de N mesuraments és la mitjana dels quadrats de les desviacions dels mesuraments al voltant de la seva mitjana μ. La variància es denota com σ_x^2 i està donada per la següent fórmula:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} x_i - \mu}{N - 1}$$

La **variància explicada** (σ_{xy}^2) és el quocient entre la variància de la predicció (σ_x^2) i la variància de la variable a predir (σ_y^2). És a dir:

$$\sigma_{xy}^2 = \frac{\sigma_x^2}{\sigma_y^2}$$

Per últim, tenim la **variància no explicada**, o residual. Que no és més que la resta la variància explicada respecte de la variància a predir.

4 Estat de l'art

L'origen dels primers fàrmacs es remunten als temps en els quals l'ésser humà utilitzava recursos naturals per la conformació de substàncies que els ajudes en diversos aspectes: combatre malalties (pal·liar el dolor o reduir els símptomes), obtenir aliment (verí per la caça), potenciar les seves relacions socials i religioses (estimulants o al·lucinògens), etc. L'evolució d'aquest fenomen ha desembocat en el fet que, avui en dia, s'hagi aconseguit l'obtenció de múltiples fàrmacs que augmenten la nostra esperança de vida i millora les condicions d'aquesta. En l'actualitat, aproximadament la meitat d'aquestes substàncies són d'origen natural, i han sigut modificades als laboratoris per potenciar la seva efectivitat i reduir els seus efectes negatius.

El desenvolupament de nous fàrmacs, és una tasca que requereix la col·laboració entre diferents professionals de diversos àmbits, a la que la capacitat de deducció, la intuïció i en moltes ocasions, la sort, juguen un paper fonamental. Cal matisar que, sense una base sòlidament ancorada en un disseny intel·ligent i racional, l'atzar resulta infructuós en aquest procés.

Al disseny de nous fàrmacs, és indispensable conèixer el procés patològic de la malaltia. D'aquesta manera, es poden definir les molècules implicades i és possible desenvolupar compostos capaços d'interactuar amb elles i que modifiquin la patologia. Per això, és necessari conèixer l'estructura tridimensional de les molècules objectiu o dianes. Els mètodes més utilitzats per obtenir aquesta estructura són la ressonància magnètica nuclear, la cristal·lografia de raigs X i els càlculs teòrics de les forces d'unió, sigui per mecànica molecular o mecànica quàntica.

4.1 Disseny de fàrmacs. Metodes tradicionals

Els mètodes tradicionals de disseny de fàrmacs [15] es basen en mètodes variacionals. Als següent apartats veurem en que consisteixen aquest mètodes.

4.1.1 Cerca del compost líder

A la cerca del compost líder, també anomenat cap de sèrie, es poden aplicar diversos mètodes. Els dos més comuns són:

- La utilització de productes actitud presents en substancies emprades a la medicina tradicional.
- L'estudi de nous compostos de la síntesi química o de la biotecnologia.

Els dos mètodes requereixen l'existència prèvia d'un ampli conjunt d'assajos biològics acuradament dissenyats, això permetrà determinar de forma ràpida i exacta l'activitat biològica dels nous compostos. En el cas de l'estudi de compostos empleats a la medicina tradicional, el disseny de la prova biològica és més fàcil, ja que existeix informació prèvia de l'activitat prevista; es tracta de confirmar de manera científica la seva eficàcia. En canvi, quan es desconeix la possible activitat, el conjunt d'assajos ha de ser prou gran, per tractar de no perdre cap informació rellevant. En el segons cas, els costos que representen aquest tipus d'estudi fan que es vegi limitat el número i la quantitat d'accions biològiques.

Altres mètodes d'obtenir un compost líder són:

- L'aïllament dels productes responsables d'una acció biològica determinada i la seva posterior identificació i caracterització. Això succeeix en el cas de productes naturals que no han sigut reconeguts per l'ésser humà.
- La detecció i observació d'efectes secundaris o accions biològiques inesperades durant l'ús de compostos actius dissenyats amb altre fi.

- L'observació del metabolisme dels compostos. A vegades, alguna de les fases del metabolisme d'un fàrmac presenta una activitat major o diferent a la del fàrmac original.
- L'anàlisi de l'activitat biològica de productes intermedis de la síntesi d'un fàrmac.

4.1.2 Desenvolupament del compost lider

Una vegada seleccionat i definit el cap de sèrie, és necessari trobar la forma de modificar la seva estructura amb l'objectiu de millorar-lo. La finalitat d'aquest procés és la de dotar als nous fàrmacs d'una major activitat, d'una millor biodisponibilitat, de menor toxicitat i d'un nombre mínim d'efectes secundaris. Al procés de trobar quines són les millors variacions que es poden aplicar a un compost se'l coneix com: Variació Molecular. Aquest procés pot tenir diferents finalitats depenent del compost i la malaltia. A continuació els expliquem:

- Millora de l'activitat del líder.
- Eliminació dels efectes secundaris no desitjats.
- Potenciació d'accions secundàries desitjades, ja sigui per una relació directa amb l'acció principal o per ella mateixa.
- Separació d'activitats en compostos amb múltiples accions. Té com a objectiu potenciar alguna de les accions farmacològiques per sobre de la resta, o eliminar algunes per potenciar a les altres.
- Combinació d'activitats. Es tracta del cas oposat a l'anterior, i el seu objectiu és reunir en un mateix compost diferents activitats que poden actuar juntes contra alguns símptomes associats.
- Modificació de la biodisponibilitat del fàrmac líder, és a dir, la modificació de la fracció i la velocitat amb què el fàrmac arriba a la seva diana terapèutica.

4.1.3 Tipus d'aplicació del mètode variacional

En aquest subapartat parlarem de dues estratègies a seguir per augmentar l'eficàcia del compost líder. És a dir, com s'ha de modificar el cap de sèrie per obtenir el resultat esperat.

La primera estratègia és la **Substitució Biosimilar**, aquesta consisteix a modificar l'estructura molecular del compost mitjançant intercanvis d'agrupacions moleculars equivalents. Aquesta es basa en la idea que si hi ha dos compostos amb una mateixa activitat biològica, també han de compartir una mateixa estructura, o almenys, punts comuns en les parts responsables de l'activitat. D'aquesta manera, es poden intercanviar grups amb la mateixa distribució per tal de modificar el comportament del fàrmac sense desestabilitzar-lo. Ara bé, no sempre les substitucions biosimilars generen una equitat respecte a l'activitat. En general, el terme biosimilitut s'aplica per a tot el conjunt d'analogies que es poden establir entre dues agrupacions atòmiques que serveixen per definir l'estructura d'una molècula.

La segona s'anomena **Modulació Molecular**, que consisteix en variar el compost líder amb modificacions. Aquestes estàn limitades per l'extensió, ja que la molècula ha de mantenir les característiques inicials, també per nombre de posicions modificades. A pesar d'aquestes limitacions, és molt freqüent trobar resultats positius. Existeixen 3 tipus generals d'actuacions diferents:

- **Modulació:** Comprèn isomerització, homologia, alquilació, ramificació, desalquilació, saturació, insaturació, canvi a la localització de la instauració, desplaçament d'una funció, introducció, substitució o eliminació d'heteroàtoms, introducció de sistemes cíclics, contracció o extensió de cicles, substitució de cicles, etc.

- **Simplificació:** A vegades, la molècula es trenca, en un intent de determinar quina o quines parts contribueixen més a l'activitat biològica que s'està estudiant. Amb aquest objectiu es dissenyen compostos més senzills, que continguin de manera aïllada les parts mencionades. Cal aclarir que les parts responsables de l'activitat no tenen per què trobar-se unides entre si, si no que poden trobar-se separades per una sèrie

d'àtoms que no formen part determinant a la interacció amb el receptor. Per la realització d'aquest tipus d'estudi és necessari un ampli coneixement de l'estructura tridimensional i del comportament conformacional.

- **Unió d'elements actius:** es tracta d'unir residus a la molècula els quals han demostrat ser actius en altres sistemes moleculars.

Amb aquest tipus d'estudi ens adonem de la importància de la combinació de diferents ciències implicades al disseny de fàrmacs. També es dedueix la necessitat de millorar les tècniques d'anàlisi instrumental, per tal de conèixer amb més precisió les estructures moleculars dels compostos. Els previsibles avanços als camps de la biotecnologia, la biologia molecular i les ciències relacionades estan canviant el disseny de nous fàrmacs.

4.2 Disseny de fàrmacs assistit per computadors

Actualment, existeixen diversos mètodes experimentals per determinar l'estructura molecular d'una substància. Les més comunes són les espectroscòpies ultraviolades (UV), infraroja (IR), la de masses, la ressonància magnètica nuclear H1 i de C13, així com les tècniques de difracció de rajos X. No obstant això, el desenvolupament assolit per la computació i la química computacional ha proporcionat la generació de sistemes que permeten calcular la geometria i l'energia molecular. Aquests sistemes són capaços de generar dades amb una àmplia aplicació a la investigació experimental, tant per la interpretació de resultats i la planificació de futurs, com per deduir informació no assequible experimentalment. Es poden considerar 3 tipus de mètodes de càlcul teòric:

- Mètodes ab initio.
- Mètodes semi-empírics.
- Mètodes de mecànica molecular.

L'elecció del mètode depèn fonamentalment de la mida de la molècula i del temps requerit. Els mètodes de mecànica molecular requereixen menys temps de càlcul respecte als de mecànica quàntica. També reprodueixen els valors experimentals referents a geometries i energies amb una bona precisió. Per això són recomanables en el cas de les macromolècules amb els quals els mètodes de mecànica quàntica presenten problemes de temps computacional. Els mètodes que relacionen estructura química amb l'activitat biològica assistits per computadors poden dividir-se en dues grans categories: SAR i QSAR. D'ells en parlem als següents apartats.

4.2.1 SAR (Structure-Activity Relationship)

Els mètodes SAR treballen amb molècules en 3 dimensions, els més importants són:

- L'anàlisi conformacional.
- La mecànica quàntica.
- Els camps de força.
- Els gràfics moleculars interactius.

Els gràfics moleculars interactius permeten la representació i la manipulació de molècules en 3 dimensions, el que dóna una informació espacial essencial per comparar molècules i estudiar la interacció entre lligands i dianes.

La cerca d'aspectes estructurals indispensables a una sèrie de molècules per aconseguir la unió al receptor i experimentar una activitat farmacològica es coneix com a **cerca del farmacòfor**, el qual és el conjunt de grups químics que totes les molècules actives sobre un mateix receptor tenen en comú i que són indispensables per la interacció amb el mateix. Actualment existeixen dues tècniques per obtenir aquest farmacòfor segons si es coneix l'estructura del receptor prèviament.

Quan no es coneix l'estructura del receptor la tasca està enfocada a la cerca de les seqüències a l'estructura química d'un conjunt de lligands que s'uneixen a un mateix receptor. Amb aquesta finalitat, s'utilitzen

softwares que permeten la visualització tridimensional de les estructures, sobre les quals es poden aplicar diferents transformacions geomètriques (rotacions, translacions, ...) i superposicions d'estructures, per detectar aquelles regions de la molècula que són essencials per a l'activitat, així com aquelles que permeten la seva substitució per variar l'afinitat. Aquest tipus d'estudi necessita una anàlisi conformacional rigorós de cadascuna de les molècules que es volen analitzar. Això és pel fet que existeix la possibilitat que, per algun fàrmac, la conformació activa no sigui la més estable termodinàmicament pel fet que l'energia lliure d'associació supera l'energia necessària perquè el lligand pateixi un canvi conformacional. Aquesta limitació fa que aquest tipus d'estudi sigui molt complicat i que els resultats no siguin sempre satisfactoris, ja que cada conformació energèticament accessible pressuposa una disposició tridimensional diferent de la d'altres grups candidats a interaccionar amb el receptor.

Quan es coneix l'estructura del receptor es simplifica l'estudi d'identificació del farmacòfor. No obstant això, el fet de determinar l'estructura d'una macromolècula, constitueix una tasca difícil. Els primers a descriure una estructura i una conformació detallada per una estructura cel·lular van ser els científics Watson i Crick l'any 1953, amb la seva investigació sobre l'ADN. Per altra banda, la primera descripció detallada sobre l'estructura i conformació d'una proteïna es va produir l'any 1958, quan Kendrew va caracteritzar per cristal·lografia de rajos X a la mioglobina de balena. Si tenim en compte que una gran quantitat de fàrmacs realitzen la seva acció per unió a enzims o receptors peptídics, ens adonem que aquest tipus d'informació estructural és de gran utilitat al disseny de fàrmacs. En l'actualitat existeixen bancs de dades d'estructures tridimensionals de macromolècules biològiques, procedents de cristal·lografia de rajos X i de ressonància magnètica nuclear. La base de dades més gran de macromolècules la constitueix el banc de dades de proteïnes (PDB [16]) que actualment conté més de 100.000 estructures de macromolècules biològiques i més de 8.000 estructures d'àcids nucleics. A més, s'ha aconseguit la cristal·lització de les estructures de receptors a la seva unió amb el lligand. D'aquesta forma, es coneix quins sons els grups del lligand que interactuen directament amb el receptor i amb quins residus de la proteïna ho fan.

Amb aquests resultats és possible analitzar la naturalesa de la unió, la flexibilitat del receptor i les interaccions que mantenen al lligand unit

al mateix. Això permet estimar l'energia d'estabilització d'aquest compost, i a més, l'aportament per separat de cadascuna de les regions del lligand. D'aquesta manera és possible definir amb precisió i exactitud el farmacòfor del fàrmac. Per altra banda, l'assimilació computacional d'aquestes dades permet la manipulació del lligand i la inserció d'altres al centre d'unió, permetent analitzar la interacció entre nous lligands i el receptor. També permet el disseny de substàncies pensades exclusivament pel seu ajust exacte amb el receptor, generant una potent font de nous candidats.

Donat que la majoria de les dades de proteïnes provenen de cristal·lografia de rajos X, la qual s'obté en estat cristal·lí, fa que el modelatge de proteïnes i receptors afronti dos problemes bàsics:

1. El gran nombre de conformacions per una proteïna. Donada la gran quantitat de graus de llibertat en qualsevol receptor provoca que la quantitat de conformacions teòriques possibles sigui molt gran. Això no passa quan es disposa de dades obtingudes per ressonància magnètica nuclear ja que ens aporta les dades dels aspectes dinàmics del moviment intern.

2. La descripció teòrica de la interacció entre els àtoms d'una proteïna requereix la formulació precisa d'un potencial o camp de força empíric que contingui termes per representar els enllaços covalents, els angles dels enllaços d'hidrogen i les interaccions electrostàtiques i de Van der Waals.

A pesar d'aquestes limitacions, és possible predir l'estructura i les propietats d'una proteïna a partir de l'estructura tridimensional d'una altre similar. D'aquesta forma s'aconsegueix una aproximació d'una proteïna no coneguda i, per tant, es poden dissenyar nous lligands capaços d'interactuar al centre actiu de la mateixa.

4.2.2 QSAR (Quantitative Structure-Sctivity Relationship)

Els estudis QSAR requereixen que l'activitat biològica entre el fàrmac i el receptor sigui definida per una funció de les característiques estructurals de la molècula. El model extratermodinàmic de Hansch dona una explicació matemàtica de la recollida en la següent equació:

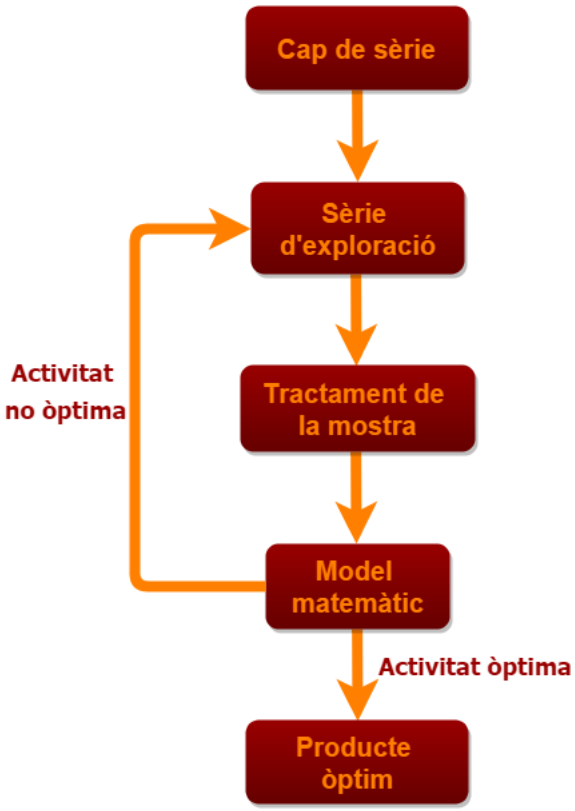
$$lnA = f_h(X_h) + f_e(X_e) + f_s(X_s) + c$$

on A és l'activitat i $f_h(X_h), f_e(X_e)$ i $f_s(X_s)$ són funcions d'índexs o paràmetres hidrofòbics, electrònics o estèrics respectivament. El terme extratermodinàmic prové del fet que les relacions es descriuen en termes termodinàmics, encara que no es dedueixi de les seves lleis. Del paradigma de Hansch [17] s'extreuen una sèrie de punts que formalitzen les assumpcions fonamentals de la metodologia extratermodinàmica:

1. L'activitat biològica és funció de l'estructura del fàrmac.
2. L'estructura del fàrmac implica certes propietats globals com la hidrofobicitat, la càrrega neta, la solubilitat, etc. i certes propietats locals com la distribució de la hidrofobicitat i la carrega i volum de determinades posicions de la molècula.
3. Aquestes propietats globals i locals poden ser quantificades mitjançant certs paràmetres.
4. Sempre existeix una funció que relaciona els canvis d'activitat biològica amb els canvis en les propietats globals i locals, tot i que potser no és senzilla ni obvia.

Les funcions que relacionen estructura i activitat, que proposa el mètode extratermodinàmic, constitueix un procediment mitjançant el qual es pot cercar els productes més actius entre un conjunt de candidats. Independentment del model que s'utilitzi, la metodologia d'investigació QSAR segueix uns passos comuns que s'il·lustren a la Figura 4.1.

Figura 4.1: Diagrama de la metodologia d'investigació QSAR.



El punt de partida sempre requereix l'existència d'un **cap de sèrie**. Aquest compost, tot i que per si sol té interès com a model a desenvolupar, té propietats que permeten modificacions amb la finalitat d'obtenir sempre un lligand el més biodisponible possible.

Quan es disposa d'un candidat, és necessari dissenyar una **sèrie d'exploració**, conformat per un conjunt d'anàlegs del candidat que han de ser sintetitzats i analitzats quant a l'activitat biològica respecta. Els membres de la sèrie d'exploració estan constituïts per un nucli comú i uns fragments variables característics de cada membre de la sèrie. Aquests fragments variables han de ser identificats per descriptors, que seran utilitzats com variables independents (X_i) al model. Els valors d'activitat biològica (A) s'utilitza com a variable dependent. La qualitat de la sèrie d'exploració és molt important, ja que condiciona en gran mesura la qualitat de les prediccions. Aquesta qualitat ve definida principalment per dos aspectes:

- *Dissimilitud de la sèrie:* Els membres de la sèrie han de ser diferents entre si respecte als paràmetres seleccionats. No té sentit intentar estudiar com influeix una característica sobre l'activitat si només s'estudien productes similars respecte a aquesta característica. A més, el rang de variació de les característiques ha de ser equiparable al que existeix a l'espai experimental.
- *Ortogonalitat de la sèrie:* Els membres han de ser seleccionats de tal forma que la variació en les característiques es produeixi de forma independent. Si la variació d'una propietat P1 sempre va acompanyada de la variació en altra propietat P2, no és possible distingir quina de les dues causa el canvi a l'activitat. En aquestes situacions es diu que P1 i P2 estan correlacionades. Una sèrie en la qual les correlacions entre els paràmetres són mínimes es diu que la sèrie és ortogonal.

Una vegada definida i avaluada l'activitat biològica de la sèrie d'experimentació es procedeix al **tractament de la mostra**. D'aquesta s'obté un **model matemàtic** que permetrà predir un o diversos compostos als quals li seran avaluades les seves activitats biològiques. Si l'activitat no és òptima, seran incorporats a la sèrie d'exploració. Aquest procés es repetirà fins que s'obtingui un compost amb l'activitat optima esperada, al qual s'anomena **producte òptim**.

En l'actualitat existeixen diferents tècniques per les quals es pot desenvolupar un estudi QSAR depenent del tipus de tractament matemàtic que se li dóna a la mostra. Els dos més importants són els tradicionals i els tridimensionals.

Els mètodes **QSAR tradicionals** inclouen el tractament estadístic de les dades per mètodes multivariats. En ells es valora l'activitat biològica com variable dependent del conjunt de descriptors moleculars que constitueixen les variables independents. Aquests mètodes multivariats són els següents:

- *Regressió:* en la que s'obté l'equació d'una recta, d'un pla o d'hiperplà segons el nombre de variables independents incloses a l'expressió. L'anàlisi de regressió múltiple és el més utilitzat dins dels QSAR tradicionals. Una vegada establerts els conjunts de valors de les variables independents X_i i l'activitat biològica A, s'obté un model amb forma d'equació d'una recta ($A = f(X_i)$), la qual descriu la dependència entre l'activitat biològica (A) en funció del conjunt de descriptors (X_i) així com la magnitud de les contribucions de cadascun d'ells.

- *Anàlisi clúster:* en aquest cas les dades, i per tant els compostos de la mostra, s'agrupen per la semblança entre ells segons els valors de similitud que se'ls exigeixi.

- *Anàlisi de components principals (PCA - Principal Component Analysis):* té com a principal objectiu la reducció de variables. En contra del que pot semblar, a la pràctica del modelatge estructural el nombre de descriptors és molt gran i la selecció dels que millor representen a la mostra a l'activitat biològica estudiada és difícil de definir a vegades. Al PCA, les dades es combinen en una línia recta. Aquesta línia és rotada de diferents formes i en cadascuna d'elles s'estableixen els coeficients de cada variable. Per tant, depenent de la rotació que es faci, els valors dels coeficients de cada descriptor estructural o físico-químic variarà i, en conseqüència, la seva influencia en aquest component. El problema per l'especialista consisteix a assignar a aquests components una significació estructural o físico-química. Recordem que cadascuna d'aquestes components és el resultat de la combinació lineal de tots els descriptors.

El model obtingut ha de ser analitzat en funció de la seva qualitat estadística, per poder avaluar la seva capacitat de predicció. Com major qualitat estadística tingui el model, més confiables i exactes seran les prediccions realitzades. La qualitat estadística d'un model s'avalua per diferents estadígrafs. Els més comunament acceptats són els següents:

- *Coeficient de regressió (r):* Com més tendeix a la unitat el valor de r, major ajust tindran les dades al model.
- *Valor de Fischer (F):* El valor F correlaciona la variància explicada pel nombre de graus de llibertat, amb la variància no explicada pel nombre de variables del model. Com més alt és el percentatge de variància explicada pel model major serà el valor de F. Mentre que l'existència de variables amb baixa contribució a l'explicació de la variància, tendiran a disminuir aquest valor, que tendeix a infinit en els millors models.
- *Desviació estàndard (s):* La desviació estàndard s depèn de la variància no explicada i dels graus de llibertat del model, i és una mesura de quant s'allunyen els valors predits pel model de la línia, pla o hiperplà. La tendència a 0 d'aquest valor podria suposar major qualitat de la predicció. En canvi, això pot conduir a models sobrepredictius en els quals es reflecteixi exactament el comportament de la mostra però que no sigui possible utilitzar-lo per a l'extrapolació dels valors en el cas de la predicció de nous compostos. Un valor de s és vàlid quan està al mateix ordre de magnitud de l'error experimental dels mesuraments de la variable dependent (activitat biològica).

Per a l'establiment d'una equació amb qualitat s'ha de considerar el nombre de variables independents que s'inclouran al model. Aquestes variables, han de ser limitades pel nombre de compostos que formen part de la sèrie d'exploració, de forma que la relació existent entre ells sigui d'una variable per cada 5 o 6 compostos a la sèrie. La utilització d'un nombre excessiu de variables en una equació de regressió pot conduir a reproduccions de la mostra d'entrenament eliminant-li el valor predictiu al model. A més, és necessari assenyalar que com major nombre de variables s'inclouen a l'equació, major serà el nombre de paràmetres a variar al compost que es vol obtenir, fet que dificulta la pràctica sintètica.

El **QSAR tridimensional** o QSAR-3D és una tècnica moderna que combina els aspectes relacionats a l'estudi QSAR tradicional amb els dels estudis SAR o de modelatge molecular. A la representació SAR dels anàlegs actius només es discrimina si aquests són actius o inactius. A la pràctica els resultats són representats de forma numèrica per reflectir les diferències quantitatives entre la unió d'un lligand o altre. En canvi, els mètodes QSAR tradicionals tracten el tema de magnituds respecte de les afinitats però en aquests és impossible considerar la diversitat química que pot interactuar amb el receptor ja que en aquest tipus d'anàlisi l'establiment d'una única sèrie amb analogia estructural és indispensable. La necessitat d'aconseguir una interfície entre ambdós tipus d'estudi requereix el desenvolupament d'un nou tipus d'anàlisi conceptual i computacional.

El primer assoliment el va aconseguir Cramer junt amb els seus companys l'any 1988 en desenvolupar l'anàlisi comparatiu de camps moleculars (CoMFA [18]). Aquest estudi defensa que un mostreig adequat dels camps elèctrics i electrostàtics al voltant d'un conjunt de molècules o fàrmacs, pot proporcionar tota la informació necessària per comprendre les activitats biològiques. Per això és necessari el càlcul dels camps de força per mecànica molecular, que només consideren les forces estèriques i electrostàtiques. Per aconseguir-ho se superposen totes les molècules en les conformacions presumiblement actives i es calculen les energies d'interacció estèrica (Van der Waals) i les electrostàtiques (Coulomb) entre cadascuna de les molècules d'interès. Posteriorment s'efectua la tècnica estadística dels mínims quadrats parcials i s'extreuen les variables que descriuen la variància del conjunt de dades, els quals se sotmeten posteriorment a una tècnica de validació creuada. L'anàlisi de regressió lineal múltiple no és aplicable en aquesta tècnica, ja que, el nombre de variables generades en aquest tipus d'estudi és molt elevat.

Els passos que componen aquest tipus de mètode són els següents:

1. Postular la conformació activa per la molècula endògena o pel fàrmac més actiu.
2. Alinear la resta de molècules i establir al seu voltant una malla tridimensional com una caixa de punts a l'espai potencial del receptor.
3. Calcular el camp que cada molècula efectuarà sobre un àtom sonda situat a cada punt de la malla.

4. Determinar una expressió lineal mitjançant la tècnica de mínims quadrats parcials que podria consistir en el conjunt mínim de punts necessaris a la malla necessaris per distingir els compostos sotmesos a examen d'acord a les activitats determinades experimentalment.

5. Realitzar la validació creuada del model (Cross-validation).

6. Ajustar l'alineament dels compostos pitjor predits i repetir els passos anteriors fins a trobar l'alineació més gran possible.

Els resultats del QSAR-3D produïts per milers de termes (als QSAR tradicionals el nombre de termes està limitat segons la mida de la mostra), es poden visualitzar gràficament en formes de mapes de contorn, poden pintarse de diversos colors per a relatar la direcció i magnitud de la interacció. D'aquesta manera les regions pintades apareixeran a aquelles regions on les diferències electròniques i estèriques produeixin una major variació de l'activitat.

4.2.3 Panorama actual i previsions futures

L'avanç sostingut que ha experimentat la bioquímica i la biologia molecular en:

- La identificació de macromolècules dianes.
- La identificació de seqüències de nucleòtids i aminoàcids.
- L'elucidació a escala atòmica de la seva estructura i del complex fàrmac-receptor.

unit als poderosos sistemes computacionals que fent ús d'aquesta informació poden crear models tridimensionals del lligand i del receptor, fan possible l'estudi de preferències configurables de la natura i les magnituds de les forces interatòmiques que governen la seva interacció així com el comportament dinàmic d'aquest complex. Aquests procediments ajuden al millor enteniment del comportament d'aquests sistemes a escala subcel·lular, fent

possible establir comparacions entre les dades teòriques i les experimentals, i fins i tot realitzar prediccions quantitatives.

Si tenim en compte els avanços que es preveu que tindran la farmacologia molecular, se suposa que el futur del disseny de fàrmacs no estigui destinat, com fins al moment, a l'obtenció de substàncies que puguin ser reconegudes pels receptors o que modulin la síntesi, el metabolisme o la recaptació dels neurotransmissors, sinó que estarà orientat a obtenir substàncies que actuïn sobre els sistemes enzimàtics activadors de la seqüència d'esdeveniments que comporta a una resposta farmacològica.

En influir mitjançant els fàrmacs sobre aquests mecanismes intermediaris entre el receptor i el compost, es pot donar origen a substàncies molt concretes que operen selectivament sobre les cèl·lules que pateixin disfunció. Enfront d'una exposició continua del compost i la molècula es produeixen fenòmens de dessensibilització o hipersensibilització que poden ser responsables de noves alteracions fisiològiques. Això ens demostra que no sempre és millor actuar sobre els receptors.

Part II

Segona part

5 Desenvolupament

A l'apartat de desenvolupament s'explica quina feina s'ha fet per a resoldre cada tasca del bloc de *Plantejament del projecte* i del bloc de *Desenvolupament del projecte*. Per facilitar l'organització d'aquesta secció, s'ha dividit l'explicació segons la tasca de la planificació a la qual correspon.

Tasques del bloc de *Plantejament del projecte*:

1. Aprofundiment en l'estat de l'art.
2. Selecció d'eines de desenvolupament.
3. Familiarització amb les dades i l'entorn.

Tasques del bloc de *Desenvolupament del projecte*:

1. Preprocés de les dades.
2. Extracció d'informació.
3. Representació de la informació.
4. Generació de recomanacions.
5. Desenvolupament d'una aplicació web.

Amb l'objectiu d'evitar la redundància d'informació, s'ha decidit no incloure la primera tasca del bloc de *Plantejament del projecte*. Aquesta

part del projecte està explicada a l'apartat 4, que explica l'*Estat de l'art* del projecte. Dit això, comencem doncs amb la *Selecció d'eines de desenvolupament.*

5.1 Selecció d'eines de desenvolupament

A l'hora d'afrontar els objectius del projecte, s'ha de fer un petit estudi de quines eines existeixen per a desenvolupar-ne una solució. D'aquesta manera podem seleccionar la que més s'ajusti a les necessitats del nostre sistema.

Recordem breument que l'objectiu del projecte és desenvolupar un software que, a partir d'una col·lecció de dades existents sobre els inhibidors de la COX-2, sigui capaç d'identificar les propietats més significatives i fer recomanacions a l'usuari de com es podrien millorar. Durant el procés, també s'han de visualitzar dades per mostrar a l'especialista la justificació de les recomanacions realitzades. Per últim, es vol proporcionar una aplicació web que faciliti l'ús del software.

Aquests objectius introdueixen una sèrie de requisits imprescindibles pel sistema:

1. *Tractament de les dades:* Llegir les dades de les quals disposem en el seu format i transformar-les, si és necessari, en el format que requereixi el nostre software.

2. *Anàlisi de dades:* Analitzar les dades per extreure informació més resumida i explicativa.

3. *Generació de gràfics:* Generar diversos gràfics que s'adaptin a la informació que es vol mostrar, facilitant la comprensió humana d'aquesta i generant una visió global del procés.

4. *Visualització molecular en 3D:* Visualitzar en 3D la interacció dels fàrmacs amb la proteïna i localitzar propietats rellevants pel procés.

5. *Computar decisions i recomanacions:* Implementar algoritmes capaços de prendre decisions encertades, a partir de la informació extreta de les dades.

6. *Automatització:* Automatitzar totes les fases del sistema de forma que es minimitzi la intervenció necessària de l'usuari.

7. *Aplicació web:* Crear una interfície gràfica en forma d'aplicació web que faciliti l'ús de les funcionalitats implementades.

Una vegada els requisits has sigut especificats, s'han de prendre decisions respecte a les tecnologies a utilitzar i com connectar-les. Per aquesta raó, volem enumerar les opcions tecnològiques que poden donar resposta a cada requisit de manera independent:

1. *Tractament de les dades:* Per llegir i transformar dades, sovint ens pot servir una eina d'ofimàtica (Excel, Calc, etc.). Tot i això, també es pot realitzar amb quasi qualsevol, per no dir qualsevol, llenguatge de programació o qualsevol eina d'anàlisi de dades (RapidMiner [19], Weka [20], ...).

2. *Anàlisi de dades:* L'anàlisi de dades es resol freqüentment amb un llenguatge de programació que facilita el desenvolupament estadístic (R [21], Python [22], Matlab [23], etc.). Una altra opció és la utilització d'eines pensades per aquest tipus de tasques concretes com les citades a l'apartat anterior.

3. *Generació de gràfics:* Per la generació de gràfics, una opció seria una eina d'ofimàtica. Per altra banda, molts llenguatges de programació disposen de llibreries que faciliten la generació de gràfics. També existeixen múltiples serveis específics per generar gràfics de tot tipus (Hohli [24], ChartGo [25], etc.).

4. *Visualització molecular en 3D:* Aquesta és una tasca més complexa que només es pot dur a terme amb eines especialitzades en aquest àmbit (PyMOL [26], Jmol [27], etc.).

5. *Computar decisions i recomanacions:* Per implementar qualsevol tipus d'algoritme mitjanament complex, la millor opció sol ser la utilització d'un llenguatge de programació.

6. *Automatització:* Tot i que no és una tasca que es pugui resoldre de forma independent, sí que és cert que l'automatització de processos es facilita amb l'ús de llenguatges de programació. En qualsevol cas, sempre és més fàcil fer-ho si la diversitat de les eines utilitzada és reduïda, ja que elimina molt soroll procedent de la comunicació entre processos.

7. *Aplicació web:* Pel desenvolupament d'una aplicació web existeixen molts frameworks, els quals et donen una base per facilitar aquest tipus de procés (Django [28], Spring [29], Shiny [30], ...). Una altra alternativa seria no utilitzar cap framework i implementar l'aplicació web des de zero, utilitzant HTML, CSS i JavaScript [31].

Analitzant l'apartat anterior, podem destacar de forma molt clara que un llenguatge de programació ens pot resoldre quasi tots els requisits. Exceptuant la visualització molecular 3D, per a la qual necessitarem una eina especialitzada, i el desenvolupament de l'aplicació web. Estudiant les opcions, ens trobem que els usuaris per als quals s'està desenvolupant aquest software estan familiaritzats amb el llenguatge de programació R, amb el visualitzador molecular 3D PyMOL i al framework per aplicacions web Shiny.

En la selecció del llenguatge de programació ens hem decantat per R principalment per tres motius: el primer és que tot l'equip el domina i es pot parlar més a baix nivell quan sigui necessari. Altre, és que com l'usuari final el coneix, pot facilitar el manteniment i l'evolució futura sense necessitar el desenvolupador actual. I per últim, R és un dels llenguatges de programació més utilitzats per a tasques similars a la nostra, ja que ofereix una gran quantitat de software que resol molt bé moltes de les necessitats que aquests impliquen.

Per a la visualització molecular en 3D hem definit PyMOL com a solució seleccionada. Aquesta solució ve donada per tres motius principals: el primer és que el desenvolupador no té experiència prèvia amb cap eina d'aquest tipus i creiem que pot ser positiu que algú del procés (l'usuari final) la conegui per suavitzar la corba d'aprenentatge. La segona és que l'equip que treballarà amb el software ja disposa de múltiples models 3D de la proteïna i els fàrmacs, i d'aquesta manera no s'ha de realitzar una transformació prèvia. Per últim, però no menys important, amb el que hem investigat sobre aquesta

hem pogut veure que és àmpliament utilitzada i que els resultats que ofereix són molt bons.

Per últim, el desenvolupament de l'aplicació web es realitzarà amb Shiny. En aquest cas, no s'ha donat tant de pes al coneixement previ del desenvolupador. La decisió final ha estat més condicionada a decisions prèvies, coneixement de l'usuari final i simplicitat del procés. Shiny és un framework capaç de convertir de forma àgil, una aplicació implementada amb R a una aplicació web. Tot i que és l'opció menys potent de les plantejades, també és cert que és la més simple i que proporciona les funcionalitats suficients per desenvolupar el producte desitjat. Això elimina quasi completament la complexitat d'adaptar software desenvolupat en R per la seva utilització des d'una aplicació web. Per altra banda, com ja hem comentat prèviament, l'usuari sí que coneix aquesta tecnologia, per tant pot proporcionar ajuda a l'aprenentatge i elimina la dependència del desenvolupador actual per realitzar el manteniment i l'evolució futura del sistema.

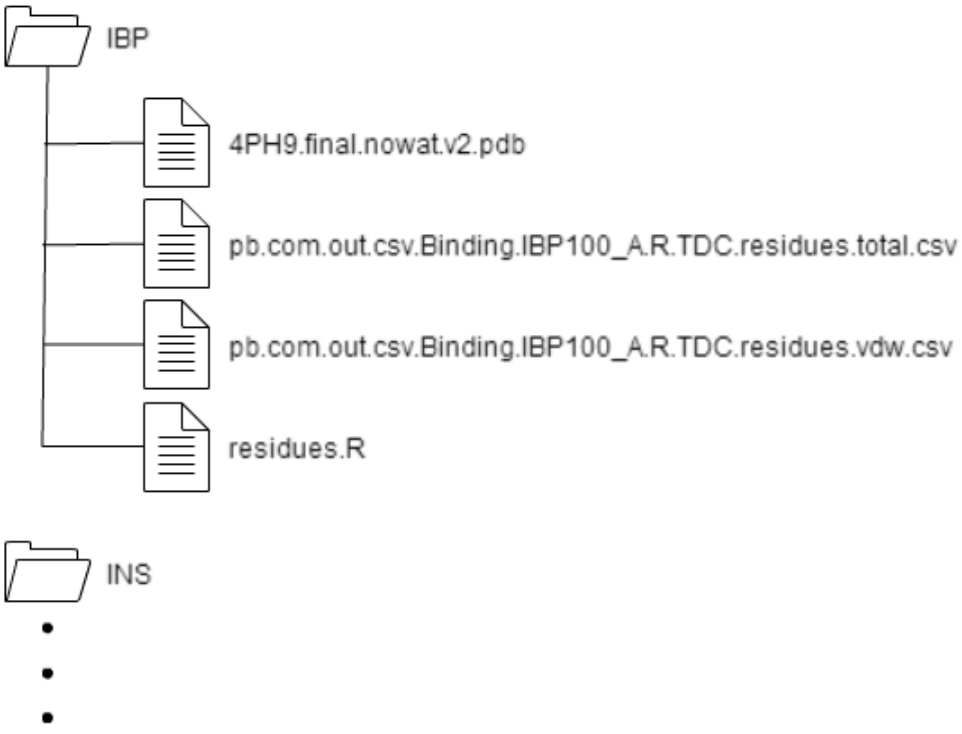
En definitiva, s'han seleccionat 3 eines per dur a terme el desenvolupament del projecte. La primera és el llenguatge de programació R, que servirà per tractar i analitzar les dades, generar gràfics, computar decisions i recomanacions i automatitzar tot el procés. La següent eina és Pymol, amb la finalitat de poder visualitzar els models moleculars en 3D. I per últim, Shiny, que facilita el desenvolupament de l'aplicació web final.

5.2 Familiarització amb les dades i l'entorn

Una vegada seleccionades les tecnologies que millor s'ajusten a les necessitats del sistema. Hem cregut adient dedicar una fase del projecte a la familiarització amb les dades i l'entorn. Amb aquesta finalitat, s'ha implementat un petit script capaç d'interpretar les dades i generar un gràfic amb elles. Això ens força a conèixer en profunditat el format de les dades i algunes funcionalitats indispensables pel desenvolupament del projecte.

En primer lloc, parlarem de les dades de les quals disposem. Un aspecte important és saber com estan organitzades. Això inclou la jerarquia de fitxers i documents que formen la col·lecció completa de dades. Podem

Figura 5.1: Exemple de la jerarquia de fitxers de les dades.



veure un petit exemple d'aquesta jerarquia a la Figura 5.1. En el nostre cas, al primer nivell disposem d'un conjunt de carpetes, on cadascuna representa la informació referent a un lligand. Dins de cadascuna d'aquests directoris trobem una sèrie de fitxers, els quals es poden classificar en 3 tipus:

- *Fitxers de dades:* Son fitxers amb extensió .csv i el nom dels quals segueixen un format concret. Aquest format, en el qual hem inclòs les parts variables entre <>, és:

  pb.com.out.csv.Binding.<Lligand>100_<Lletra>
  .R.TDC.residues.<Tipus>.csv

  Aquestes parts variables tenen el següent significat:

  - *Lligand:* Es tracta de 3 caràcters alfanumèrics que defineixen quin és el lligand al qual pertanyen les dades. Aquests 3 caràcters coincideixen exactament amb el nom de la carpeta que conté el fitxer. Al desenvolupament d'aquest projecte, tenim un conjunt de 5 lligands (INS, KES, MTZ, IBS, IBP).
  - *Lletra:* Es tracta d'una lletra majúscula que determina quina és la cavitat de la proteïna a la qual fan referencia. En el nostre cas, aquesta lletra només pot ser A o B, ja que la COX-2 només disposa de 2 cavitats. Tot i això, en aquest projecte sempre treballarem amb la cavitat A.
  - *Tipus:* Es tracta d'una paraula d'entre 3 i 5 lletres en minúscules que identifica l'estudi de quin tipus d'energia ha proporcionat aquestes dades. Aquestes paraules coincideixen amb les abreviatures dels 10 tipus d'energia explicats a l'apartat 3.1 del document.

- *Fitxers de models:* Son fitxers amb extensió .pdb que contenen un model molecular tridimensional. Aquests models són interpretables per simuladors com el PyMol.

- *Altres:* Son fitxers que no compleixen les característiques dels fitxers de dades ni les dels fitxers de models. No s'explica res més d'ells, ja que no són rellevants per al nostre projecte.

Amb l'estructura de fitxers definida, podem parlar de quin format tenen els fitxers i quin és el seu significat. Els fitxers de dades són matrius numèriques amb 1107 columnes i 82 files. Cada fila representa un instant de temps, el qual compon la primera columna de la matriu anomenada "time". A causa d'un problema de precisió, els 82 instants estan representats per 82 números d'un decimal entre el 0.0 i el 1.0 (ambdós inclosos), el que provoca una repetició de nombres. Les altres 1106 columnes representen els 1105 aminoàcids de la proteïna i el mateix lligand. Els noms d'aquestes columnes estan formats per 3 lletres majúscules, que determina el tipus del residu, seguides d'un número, que identifica el residu en qüestió. En el cas de la columna que correspon al lligand, segueix el mateix format, on les 3 lletres corresponen al nom del directori, i el numero és l'identificador al model molecular. D'aquesta forma, el número de cada cel·la, representa la contribució a l'energia d'unió de l'àtom identificat amb el nom de la columna amb el lligand a l'instant de la columna "time". Cal remarcar que aquesta energia fa

Figura 5.2: Fragment d'una de les matrius de dades en el format original.

	time	ACE1	HID2	HIE3	PRO4	CYX5	CYX6
1	0.0	0.00000000	0.06666667	0.00000000	0.00000000	-0.06666667	-0.06666667
2	0.0	0.00000000	-0.03333333	0.03333333	0.03333333	0.03333333	-0.06666667
3	0.0	-0.10000000	0.00000000	0.03333333	0.00000000	0.03333333	0.03333333
4	0.0	0.00000000	0.03333333	0.03333333	0.03333333	0.03333333	0.00000000
5	0.0	0.00000000	0.00000000	0.03333333	0.03333333	-0.06666667	0.03333333
6	0.1	0.00000000	-0.03333333	0.03333333	0.00000000	0.03333333	0.03333333
7	0.1	0.00000000	0.10000000	0.00000000	0.03333333	0.00000000	0.00000000
8	0.1	0.00000000	0.00000000	0.00000000	0.03333333	0.03333333	0.03333333
9	0.1	0.00000000	0.00000000	0.03333333	0.00000000	0.03333333	0.03333333
10	0.1	0.00000000	-0.03333333	0.03333333	0.00000000	0.00000000	0.03333333

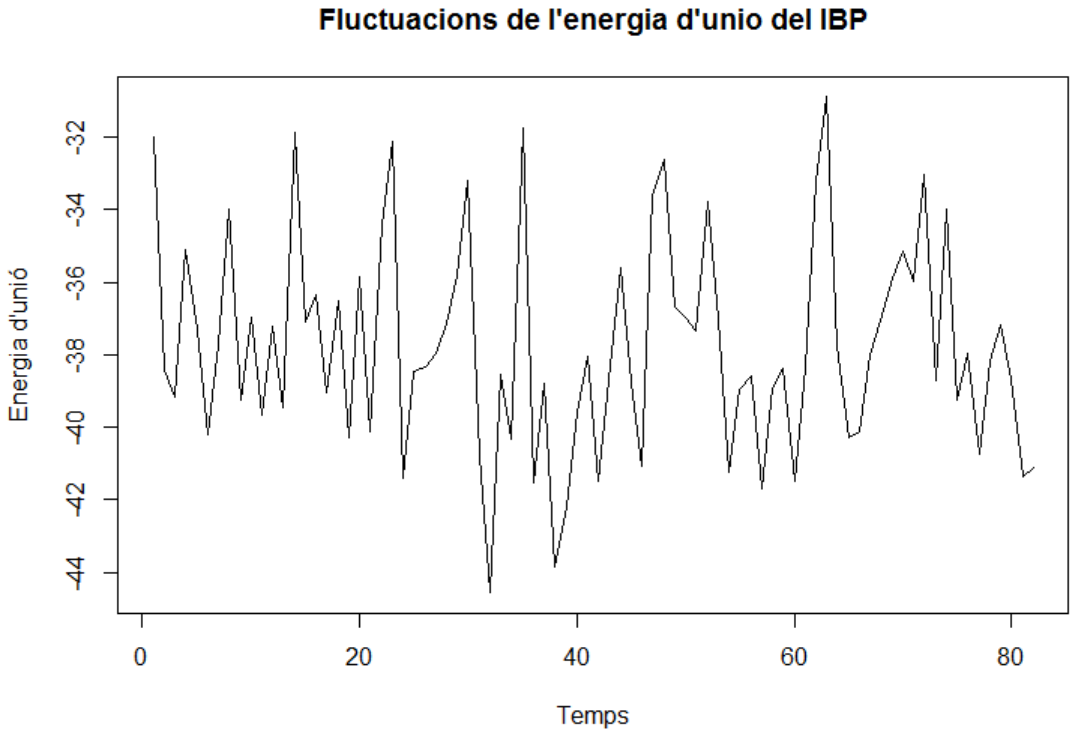
referencia sempre al tipus d'energia que apareix al nom del fitxer. Per acabar de comprendre el format de les matrius, a la Figura 5.2 tenim el fragment d'una d'elles.

Per fer una petita prova de concepte, hem decidit mostrar la fluctuació de l'energia total d'unió de cada lligand al llarg del temps. Amb aquesta finalitat, mostrarem una gràfica on l'eix Y representarà l'energia d'unió i l'eix X el temps. A la gràfica es mostrarà una línia que representarà aquesta fluctuació. Com que volem mostrar l'energia total, haurem d'utilitzar els arxius amb tipus d'energia total. A més, com disposem de l'energia separada per aminoàcids, farem la suma de tots. A la Figura 5.3 podem veure la gràfica generada pel IBP.

5.3 Preprocés de les dades

En aquest subapartat explicarem el preprocés que s'ha realitzat per adaptar les dades originals a les necessitats del projecte. Les finalitats d'aquesta fase són eliminar dades que no aporten informació, corregir errors quan sigui possible i detectar els que no es poden corregir per tal de no donar

Figura 5.3: Gràfica de les fluctuacions de l'energia d'unió del IBP al llarg del temps.



per bons resultats erronis.

Abans d'explicar les mesures que s'han pres, s'ha d'aclarir que el nostre procediment de preprocés crea un nou directori on s'introdueix la informació que utilitzarà el sistema. Tenint en compte això, quan parlem a apartats posteriors d'eliminar arxius o modificar característiques d'aquests, el que realment estem fent és no copiar-los o modificar la còpia que es realitza. D'aquesta forma, evitem conseqüències irreversibles en el cas d'errors d'implementació.

Per començar, s'eliminen tots els fitxers que no contenen dades d'energia. Tot i que a l'apartat anterior hem dit que teníem fitxers amb models moleculars i fitxers d'altres tipus a la mateixa carpeta. Considerem que aquests no són rellevants per a l'anàlisi de les dades. De fet, més endavant sí que farem servir els models moleculars per calcular distancies entre àtoms, però demanarem a l'usuari que especifiqui sobre quin model vol que es realitzin les mesures.

En segon lloc, el que fem és eliminar les matrius que corresponen al tipus d'energia *int*. Com ja s'explica a l'apartat 3.1, aquest tipus d'energia mai contribueix a l'energia d'unió del lligand i, per tant, no aporta cap informació al sistema. Aquesta mesura té com a principal objectiu agilitzar el processament posterior de les dades.

El següent que s'ha tingut en compte és l'eliminació de la columna de les matrius que contenen la informació del lligand. Aquesta dada no és útil, ja que representa l'energia d'unió total, la qual podem calcular a partir de la seva descomposició per cada àtom, que està representada a la resta de columnes. Hem detectat que mantenint emmagatzemada aquesta dada a la matriu, hem d'anar amb cura de no utilitzar-la mai per no distorsionar els resultats obtinguts. Donat aquest inconvenient i la possibilitat de calcular-la al cas de necessitar-la, s'ha considerat que la millor solució és eliminar-la. Aquest canvi no és computacionalment significatiu, però evita una possible font d'errors a la implementació.

A continuació, s'ha tingut en compte la columna que conté l'instant de cada mostra. Aquesta columna anomenada "time", com ja s'ha explicat a l'apartat anterior, no és del tot correcta per una raó de precisió. En primer

lloc, s'ha considerat l'opció de substituir els instants, alguns dels quals es repeteixen, per una seqüència de números consecutius i diferents per cada fila. Aquesta és una opció possible, ja que en cap cas s'analitza d'alguna manera la diferència entre un instant i altres, si no que només s'utilitza com a mesura d'ordenació temporal de les dades. Posteriorment, s'ha detectat que per simplificar més la matriu, es pot suprimir aquesta columna i utilitzar com instant de la mostra l'índex de la fila a la matriu. Aquests índexs representen la mateixa seqüència de números, i com en el cas de la columna del lligand, és una manera d'evitar una possible font d'errors a l'anàlisi. Ja que si es manté la columna de temps, s'ha d'anar amb cura de no contemplar aquests valors en calcular sumes i mitjanes de l'energia dels diferents aminoàcids.

Un error, per al qual s'ha contemplat que el sistema doni una solució, és el nomenament erroni dels fitxers respecte del nom del directori que el conté. Aquest cas es dóna quan, el directori que conté la informació d'un lligand concret i que té com a nom l'abreviació que l'identifica, conté algun fitxer de dades que, al lloc on hauria de ser-hi el nom del lligand hi ha un nom diferent del nom del directori. Per solucionar aquest fet, s'ha considerat que si un fitxer és dins del directori pertanyent a un lligand, aquest fitxer conté dades del lligand en qüestió i per tant es realitza una correcció al seu nom. Posteriorment, s'ha detectat que això és perillós, ja que pot fer indetectable un error humà. Per aquesta raó, s'ha decidit la realització d'una tasca exclusivament informativa perquè l'usuari s'assabenti de la situació.

Per últim, en aquest procediment s'identifica el fet que la mida de les matrius no coincideixi amb la resta. No s'ha establert cap correcció per aquest fet i l'únic que es realitza és la comunicació d'aquest a l'usuari. El fet de que no es realitzi cap correcció és perquè la causa d'aquest pot venir donada per múltiples factors, la identificació dels quals ha de realitzar-la l'usuari.

5.4 Extracció d'informació

L'objectiu principal del bloc de feina d'*Extracció d'informació*, és el de generar un conjunt de matrius que reflecteixin informació d'utilitat a partir de les dades inicials. A continuació s'expliquen aquestes matrius.

Figura 5.4: Matriu amb la contribució de cada lligand per cada tipus d'energia.

	total	vdw	ele	cal	sur	pol	npol	gas	sol
INS	-23.60854	-26.16829	-3.476829	7.973171	-1.951220	4.496341	-28.10488	-29.64512	5.919512
MTZ	-20.44146	-20.92439	-3.629268	5.531707	-1.606098	1.902439	-22.34390	-24.55366	4.153659
IBP	-20.27154	-15.07561	-13.681301	9.830488	-1.389024	-3.850813	-16.42073	-28.75691	8.343902
KES	-19.10447	-20.62073	-3.909350	6.965854	-1.606098	3.056504	-22.16098	-24.53008	5.484146
IBS	-16.06463	-16.81341	-4.034146	6.019512	-1.317073	1.985366	-18.05000	-20.84756	4.629268

La primera matriu, conté la contribució de cada lligand a l'energia d'unió per cada tipus. Dit d'una altra manera, volem representar tota la informació d'un fitxer de dades original en un únic número que quantifiqui l'energia d'unió d'aquest lligand per aquest tipus d'energia en concret. D'aquesta manera, les columnes de la matriu resultant representen els tipus d'energia, i les files, els lligands. Pel càlcul de cada valor, primer es calcula la mitjana de la contribució de cada residu al llarg del temps. D'aquesta manera s'obté un vector on cada posició representa la mitjana de la contribució de cada residu. Finalment es realitza la suma de tots aquests valors, el que dóna com a resultat la contribució total del lligand. Podem veure un exemple del resultat final d'aquesta matriu a la Figura 5.4.

A continuació, generem la matriu amb la informació de la correlació entre l'energia de cada lligand per cada tipus amb l'energia total d'aquest lligand. Això ens permet deduir quin tipus d'energia està més correlacionat amb l'energia total. La matriu resultant tindrà com a columnes els tipus d'energia, i com a files, els lligands. Per calcular la correlació, primer obtenim un vector amb la suma de totes les columnes, tant pel tipus especificat com per al total. D'aquesta manera tenim dos vectors, el primer representa la contribució total del lligand en un instant de temps i, el segon representa la contribució del tipus energètic en qüestió en un instant de temps. D'aquesta manera, el resultat de cada cel·la serà la correlació de Pearson entre aquests dos vectors. S'ha de notar que quan es realitzi el càlcul per a l'energia de tipus total, la correlació haurà de ser 1 per què es compara amb si mateixa. Podem veure un exemple del resultat final d'aquesta matriu a la Figura 5.5.

La tercera matriu, ha de representar la desviació de la contribució

Figura 5.5: Matriu amb la correlació de Pearson de cada lligand per cada tipus d'energia respecte a l'energia total del lligand.

	total	vdw	ele	cal	sur	pol	npol	gas	sol
INS	1	0.7266393	-0.17787442	0.6955488	0.0002349027	0.6720524	0.7431368	0.5728141	0.38855135
MTZ	1	0.5943275	0.05637425	0.6054991	-0.0526827427	0.6294347	0.6854125	0.5667263	0.05665490
IBP	1	0.5567350	0.30364752	0.3466548	-0.4361717086	0.7467063	0.5068808	0.5763859	0.08454996
KES	1	0.7014021	0.04733510	0.6006140	-0.1577992775	0.6976163	0.7430372	0.6197930	0.25435010
IBS	1	0.6019806	0.29301418	0.4140510	-0.2251528532	0.7026880	0.6908573	0.6576973	-0.05958063

Figura 5.6: Matriu amb la desviació estàndard de cada lligand per cada tipus d'energia.

	total	vdw	ele	cal	sur	pol	npol	gas	sol
INS	1.790394	1.201119	0.8726871	1.3807360	0.2195640	1.1980323	1.3258030	1.252677	1.296285
MTZ	1.297571	1.021686	0.7905024	0.9175757	0.2343088	0.9534830	1.0175130	1.150079	1.155256
IBP	1.756674	1.135680	1.8310843	1.7216622	0.1879099	1.5438824	1.1912995	2.061599	1.592118
KES	1.747558	1.192853	0.9793656	1.2822838	0.2369287	1.1704370	1.2530216	1.424714	1.134256
IBS	1.377263	1.072065	0.9408133	1.0248082	0.1676305	0.9961668	0.9803313	1.400392	1.315669

energètica de cada lligand per cada tipus d'energia. Aquesta informació ens descriu l'estabilitat de l'energia d'unió del lligand, fent visibles la mida de les fluctuacions d'aquesta. El resultat ha de ser una matriu on les columnes representaran els diferents tipus d'energia i, les files, representaran cada lligand. En primer lloc, es calcula la suma de totes les columnes, el que produeix un vector on cada posició és la contribució energètica del tipus especificat del lligand en un instant de temps. El valor que introduirem a la matriu serà la desviació estàndard d'aquest vector de contribucions. Podem veure un exemple del resultat final d'aquesta matriu a la Figura 5.6.

La següent matriu, en realitat consisteix a construir 8 d'elles. Es generarà una matriu per cada tipus d'energia, excepte per a l'energia total. En aquesta, es vol visualitzar la contribució de cada residu de cada lligand pel tipus energètic en curs. D'aquesta manera, les columnes de la matriu seran els residus del lligand i, les files, els lligands analitzats. El càlcul de cada cella consisteix a fer la mitjana dels valors de la contribució del residu, al lligand especificat i pel tipus pertanyent a la matriu, al llarg del temps.

Figura 5.7: Fragment de la matriu amb la contribució de cada residu de cada lligand per un tipus d'energia.

	LYS48	LEU49	LEU50	LEU51	LYS52	PRO53
INS	0.000000000	-0.001219512	0.000000000	-0.001219512	-0.003658537	0.000000000
IBP	0.000000000	0.000000000	-0.001219512	0.000000000	-0.003658537	-0.001219512
MTZ	0.000000000	-0.001219512	-0.001219512	0.000000000	-0.003658537	-0.003658537
KES	-0.002439024	-0.001219512	0.000000000	0.000000000	-0.001219512	-0.002439024
IBS	-0.002439024	0.000000000	-0.001219512	0.000000000	0.000000000	-0.001219512

Figura 5.8: Fragment de la matriu amb la correlació de Pearson de cada residu de cada lligand per un tipus d'energia, respecte a l'energia total del lligand.

	LYS48	LEU49	LEU50	LEU51	LYS52	PRO53
INS	0.00000000	0.022362183	0.000000000	-0.1170905	-0.07029266	0.000000000
IBP	0.00000000	0.000000000	-0.004461362	0.0000000	0.05343471	0.004730389
MTZ	0.00000000	-0.003572364	-0.147167383	0.0000000	-0.05999174	-0.075102291
KES	-0.07778583	-0.008815663	0.000000000	0.0000000	-0.07918494	0.042075401
IBS	0.08879200	0.000000000	0.065102381	0.0000000	0.00000000	-0.035009441

Podem veure un fragment del resultat final d'un exemple d'aquesta matriu a la Figura 5.7.

Per últim, tenim un cas molt semblant a l'anterior. L'única diferència és que en aquesta ocasió volem mostrar la correlació de la contribució de cada residu de cada lligand amb l'energia total del lligand. Per obtenir aquest resultat, s'ha dissenyat una matriu on les columnes són els residus i, les files, són els lligands seleccionats. El resultat és la correlació de Pearson entre dos vectors, el primer és la contribució energètica de la tipologia tractada d'un residu del lligand al llarg del temps, l'altre és la contribució energètica total del lligand al llarg del temps. Podem veure un fragment del resultat final d'un exemple d'aquesta matriu a la Figura 5.8.

Aquestes matrius generades a partir de les dades originals, ens aporten una informació que les dades originals no donen sense un tracta-

ment previ. Per aquesta fase del projecte s'ha d'entendre molt bé l'àmbit de les dades per tal de conèixer quina informació pot ser rellevant, i el significat de cada valor extret.

5.5 Representació de la informació

A l'apartat de *Representació de la informació*, es visualitza la informació extreta en un format agradable i comprensible. Això facilita la feina dels usuaris, que amb un cop d'ull poden absorbir tota la informació sense necessitat de recórrer i analitzar matrius de dades numèriques. Aquesta tasca sol realitzar-se mitjançant gràfics. Tot i que una de les parts més importants de la construcció d'aquests gràfics és disposar de les dades que es volen representar, l'elecció del tipus de gràfica i la presa d'algunes decisions de disseny poden ser determinants per obtenir una gràfica satisfactòria. A continuació, descriurem les 3 gràfiques diferents que s'han desenvolupat al projecte, explicant quin procés s'ha seguit i les decisions preses.

5.5.1 Perfils globals per lligand

La primera gràfica, pretén mostrar els perfils energètics globals per cada lligand. Anomenem perfils energètics globals a la divisió de la contribució energètica d'unió total del lligand per cada tipus energètic. S'ha considerat que la realització d'un gràfic de barres pot ajudar a la comprensió del mateix. De fet, per poder mostrar la informació de tots els tipus i tots els lligands, es realitzarà un gràfic de barres agrupades. Per altra banda, s'ha considerat idoni utilitzar un degradat de colors des de blau fins a vermell, on el blau és el lligand amb major contribució total i el vermell el que menor contribució aporta. Podem veure l'exemple d'una primera versió a la Figura 5.9.

En aquesta primera versió, podem detectar que el resultat és una mica confús, donat que es representen un nombre alt de dades, inconvenient que es podria veure accentuada si s'augmentés la quantitat de lligands inclosos. A banda d'això, donat que el total no és la suma de la resta, la gràfica

Figura 5.9: Gràfica de perfils energètics globals per lligand (Versió 1).

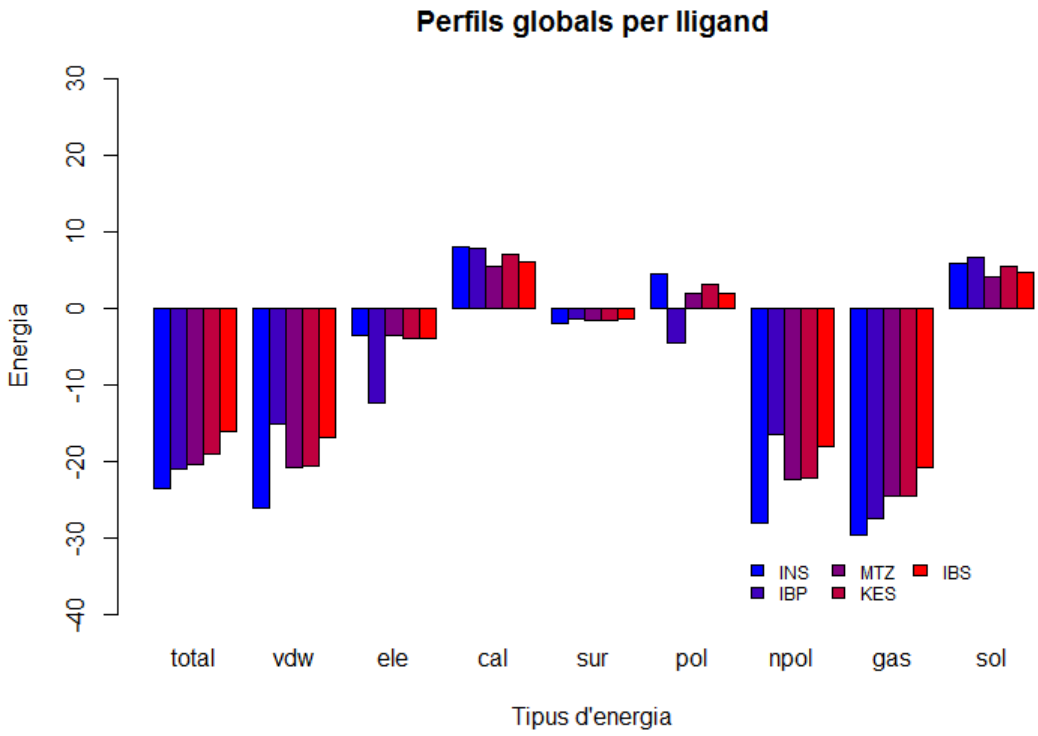
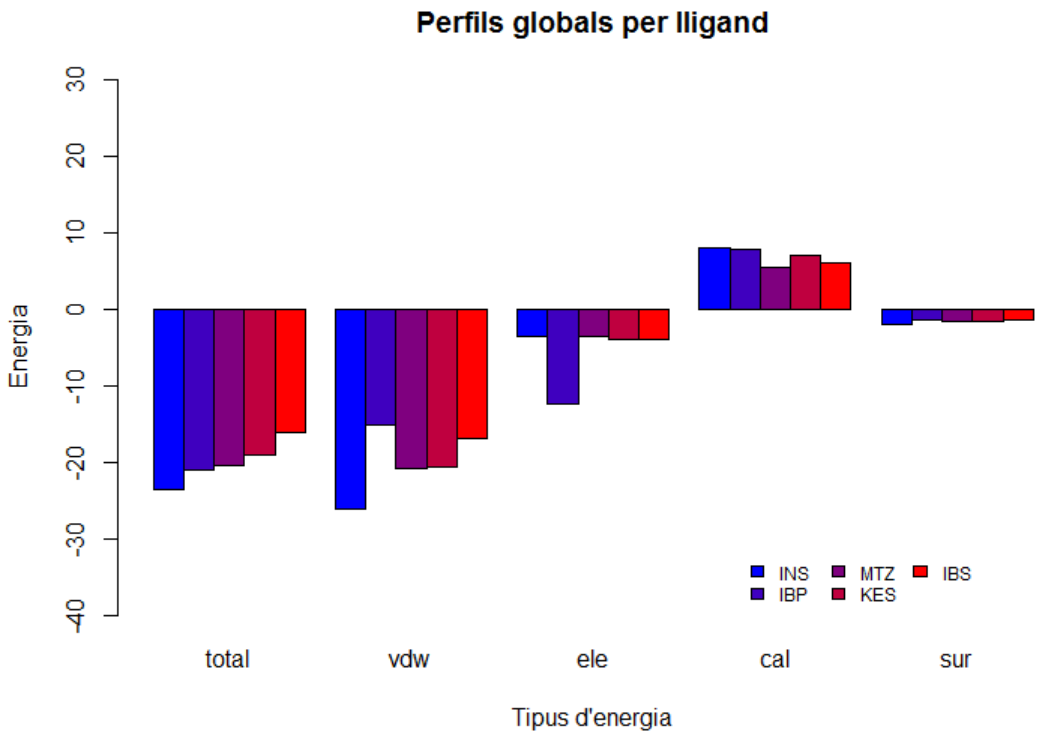


Figura 5.10: Gràfica de perfils energètics globals per lligand (Versió 2).



és antiintuïtiva. Per tractar de mitigar aquests efectes, s'ha decidit dividir aquesta gràfica en 3, on cadascuna mostrà el total i una de les descomposicions de l'energia total explicades a l'apartat 3.1. A la Figura 5.10 es mostra l'exemple per a la primera descomposició.

Aplicant aquesta millora, s'obté una gràfica més comprensible on el resultat de l'energia total és la suma de la resta d'energies a les 3 gràfiques. A partir d'aquesta versió, volem afegir la desviació de les dades que es mostren i quina correlació tenen amb l'energia total. Amb aquesta finalitat, realitzarem una nova gràfica representant les desviacions amb una línia a cada barra la mida de la qual augmentarà en funció de la desviació. Per mostrar la correlació, s'afegirà una esfera a l'extrem de la línia de la desviació seguint el

Figura 5.11: Gràfica de perfils energètics globals per lligand (Versió 3).

mateix criteri, el radi de l'esfera augmentarà en funció de la correlació amb l'energia total. També s'ha decidit reduir els espais inferiors i superiors per tal d'ampliar la zona útil de la gràfica. El resultat es mostra a la Figura 5.11.

Tot i que la gràfica resultant sembla que assoleix el seu objectiu, hi ha un aspecte que no resol. Amb aquest sistema d'esferes, no podem mostrar les correlacions negatives. Al nostre cas, les correlacions negatives també poden ser d'utilitzar per detectar energies contraproduents al fàrmac. Per resoldre aquest aspecte, es plantegen 3 alternatives que discutim a continuació:

- Correlacions en valor absolut: La primera opció consisteix a mostrar

les correlacions en valor absolut. Aquesta tècnica resol el fet que no es mostrin les correlacions negatives, però a la vegada, introdueix l'impossibilitat d'identificar quan és una correlació positiva i quan és una negativa.

- Identificar el signe segons l'estil del cercle: Sense deixar la idea de mostrar les correlacions en valor absolut, es proposa identificar el signe segons el cercle de l'esfera. D'aquesta manera es podria mostrar un cercle uniforme per correlacions positives i un cercle de puntets per a correlacions negatives. Aquesta opció resol la situació, però es considera que el resultat no és prou explicatiu, ja que és difícil detectar l'estil del cercle amb esferes de dimensions reduïdes.
- Utilitzar triangles: L'ultima i definitiva, consisteix a substituir les esferes per triangles equilàters, de manera que identificarem el signe de la correlació segons l'orientació d'aquest. Mostrant correlacions positives amb triangles drets i correlacions negatives amb triangles cap per baix.

Una vegada identificada una possible solució pel problema de les correlacions negatives, sorgeix el fet que es troba a faltar algun element explicatiu que descrigui la interpretació que s'ha de fer de les línies i els triangles. Amb aquest objectiu, hem tornat a afegir els marges superiors per tal de dibuixar una llegenda amb els valors màxims de les desviacions i les correlacions i quina figura resulta d'aquest valor. Podem veure l'exemple de la versió resultant a la Figura 5.12.

La gràfica obtinguda es considera apropiada per a la informació que es pretenia mostrar. Aquesta és capaç de transmetre les contribucions de cada tipus, la seva correlació amb l'energia total i la seva desviació. Tot això d'una forma suficientment auto explicativa i atractiva visualment. Als annexos 10.1, 10.2 i 10.3 es poden veure les 3 gràfiques resultants.

5.5.2 Perfils energètics per residu

Per continuar, es vol dissenyar una segona gràfica, capaç de mostrar els perfils energètics pels residus més importants. A diferència dels perfils globals, aquest gràfic identifica la contribució dels 10 residus que es consideren

Figura 5.12: Gràfica de perfils energètics globals per lligand (Versió 4).

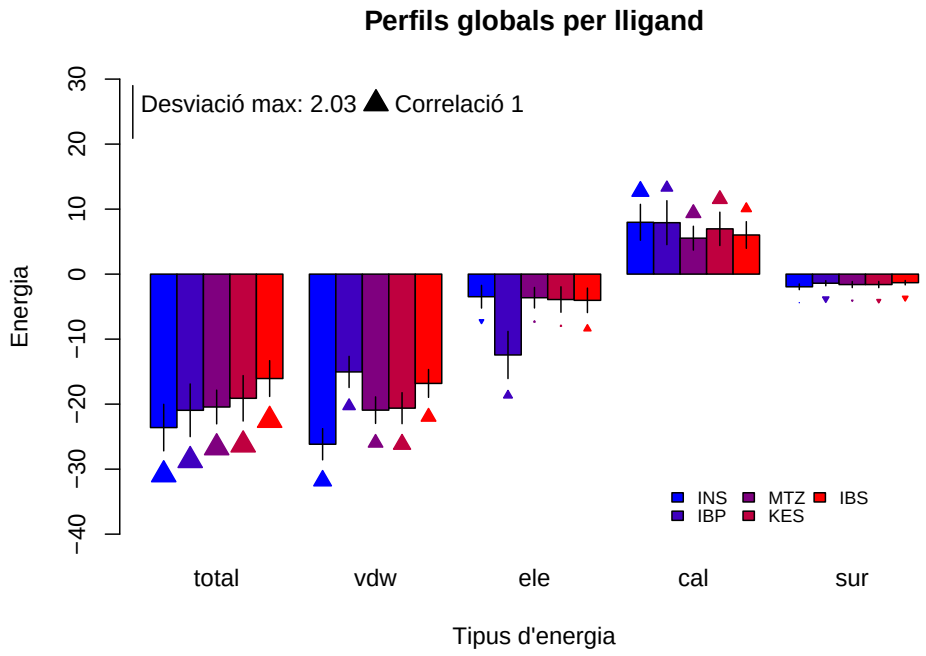
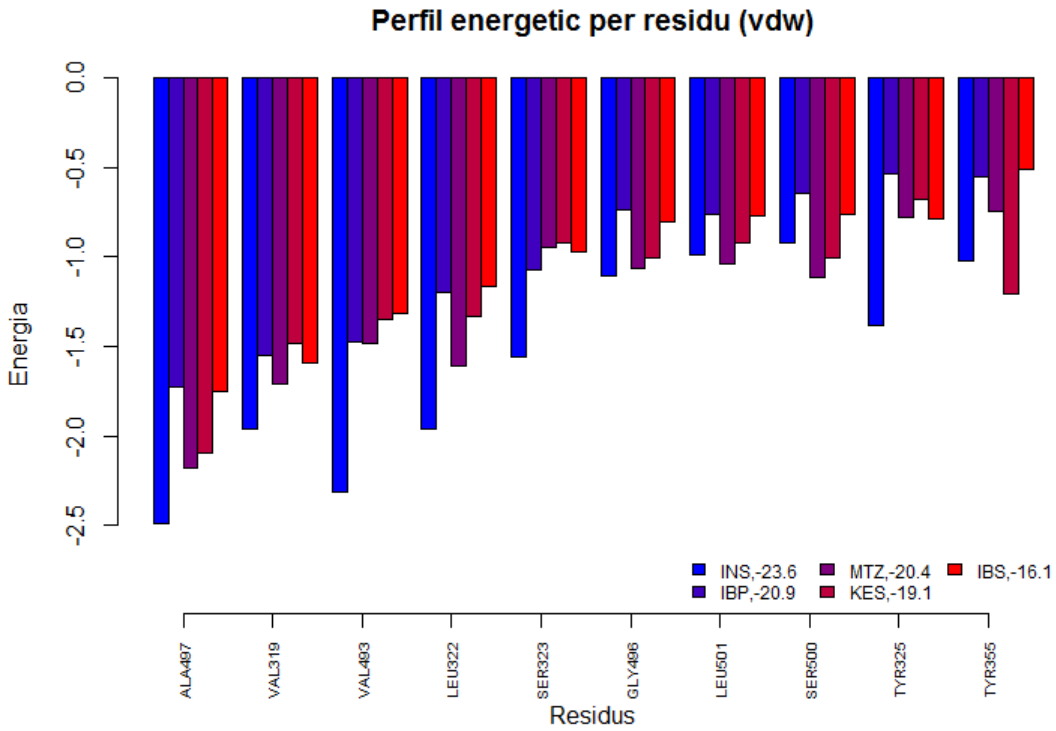


Figura 5.13: Gràfica de barres de perfils energètics per residu (Versió 1).



més importants en funció de la seva contribució energètica. A més, s'ha decidit dividir la informació segons el tipus d'energia exceptuant el total, originant la construcció de 8 gràfics. En aquest cas, no tenim clar quin tipus de gràfic pot encaixar més amb la informació que es vol transmetre. S'estudien dues possibilitats: gràfic de barres o gràfic de línies. En canvi, s'ha mantingut la decisió d'utilitzar un degradat de colors des de blau fins a vermell, on el blau és el lligand amb major contribució total i el vermell el que menor contribució aporta. Per no saturar el document amb figures, s'utilitzarà com a exemple les gràfiques pel tipus energètic de Van der Waals. Podem veure un primer exemple per cada tipus de gràfic a les Figures 5.13 i 5.14.

Figura 5.14: Gràfica de línies de perfils energètics per residu (Versió 1).

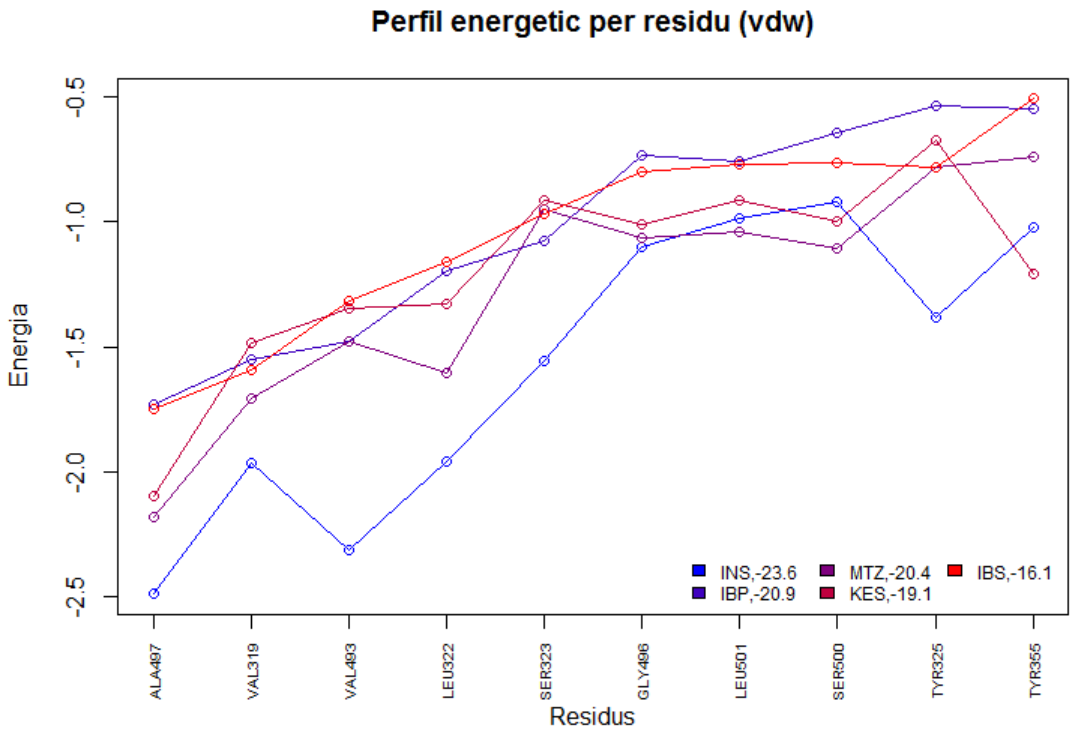
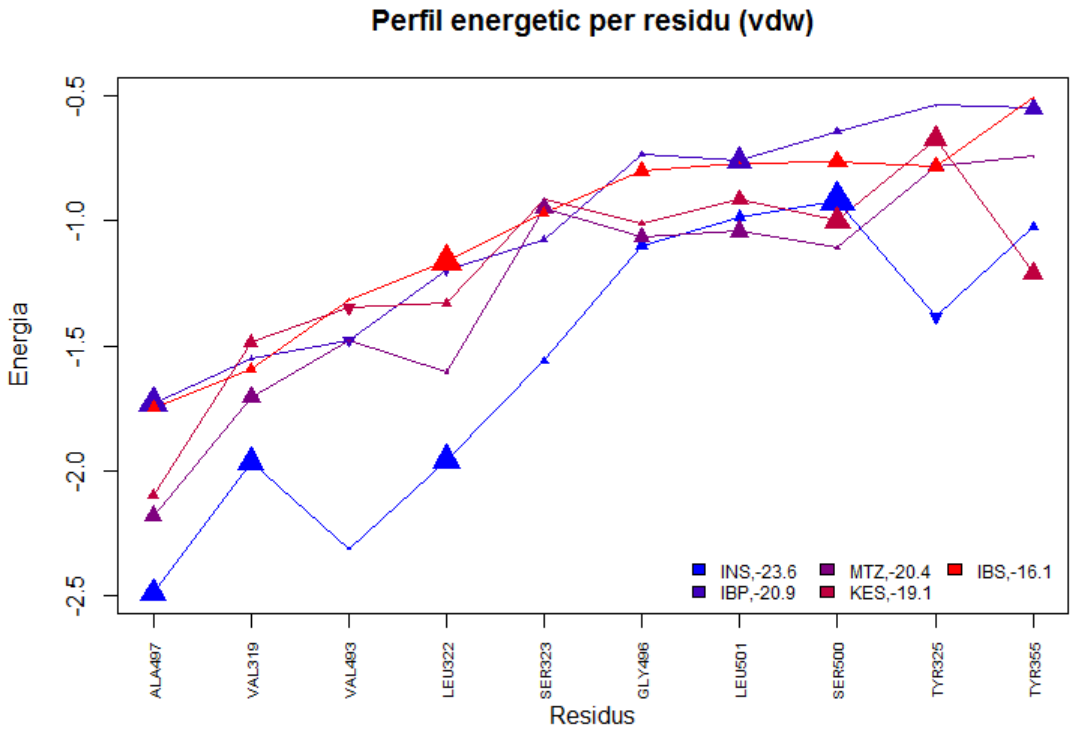


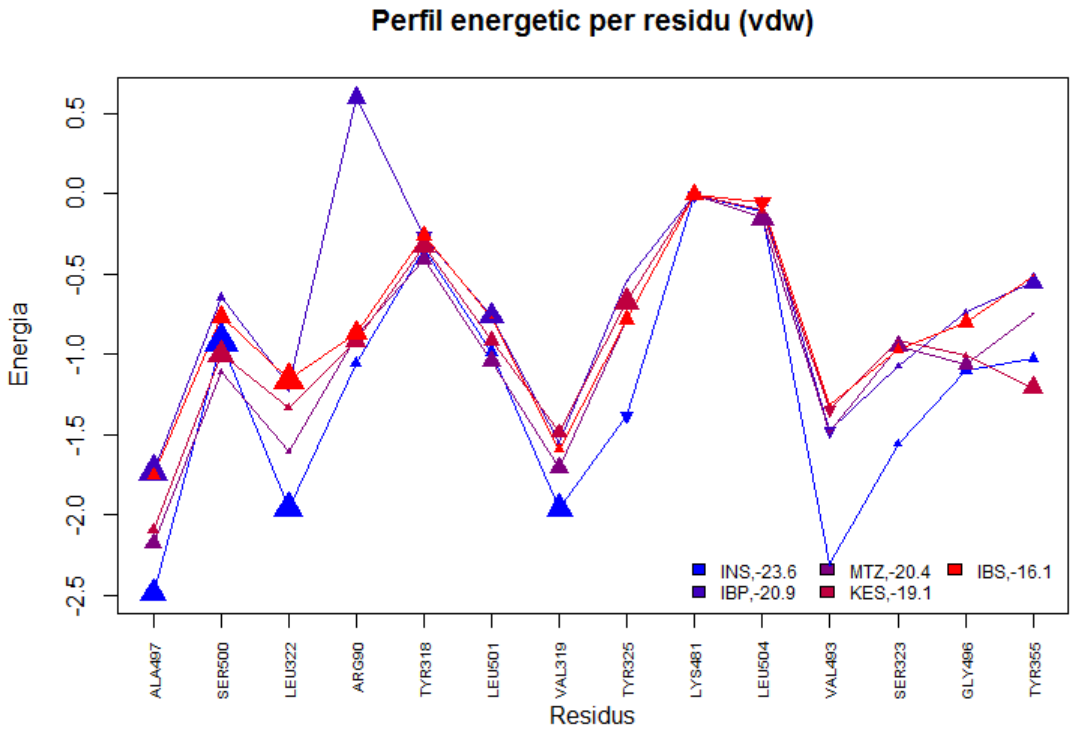
Figura 5.15: Gràfica de línies de perfils energètics per residu (Versió 2).



Analitzant els resultats, decidim que és més fàcil d'interpretar el gràfic de línies que el de barres. A banda d'això, es decideix incorporar la correlació de l'energia de cada residu amb l'energia total. Per incloure aquesta informació, utilitzarem el mètode emprat a la gràfica anterior, en la qual s'han afegit triangles amb mida variable segons la correlació. A més, el triangle estarà apuntant cap a dalt o cap a baix segon si la correlació és positiva o negativa. El resultat es pot veure a la Figura 5.15.

En incloure les correlacions de les contribucions de cada residu amb l'energia total, ens adonem que no té gaire sentit seleccionar els residus només en funció de la seva correlació amb l'energia total, sinó que hauríem de tenir en compte també la seva correlació. Per donar importància als residus amb

Figura 5.16: Gràfica de línies de perfils energètics per residu (Versió 3).



una contribució alta en valor absolut (per incloure també els de correlacions molt baixes), s'ha decidit canviar el mètode pel qual se seleccionen els residus que es mostren. De la mateixa manera que s'agafen els 10 residus amb majors contribucions, ara també s'agafen els 10 residus amb correlacions absolutes més altes. El conjunt de residus que s'inclourà a la gràfica serà la unió del conjunt de residus més importants contributivament amb el conjunt de residus més importants correlacionalment. Podem veure l'exemple de la versió resultant a la Figura 5.16.

La gràfica resultant dona una bona representació de la informació que es vol mostrar. Aquesta pretén ser una idea de quins residus són més importants, tant per contribucions altes com per correlacions altes, i comparar-

los entre ells respecte la seva aparició en els diferents lligands. Als annexos entre el 10.4 i el 10.11 (ambdós inclosos) es poden veure les 8 gràfiques resultants.

5.5.3 Correlacions totals entre tipus d'energia i energia total

Per finalitzar, s'ha implementat una tercera gràfica molt senzilla que mostra la correlació dels tipus energètics amb l'energia total. A diferència de la informació inclosa a la gràfica de perfils energètics globals per lligand, en aquest cas no es descompon la contribució entre els diferents lligands, sinó que s'analitza la correlació de la suma de tots els lligands. Per aquesta gràfica, s'ha decidit utilitzar una col·lecció de colors el més diferent entre ells possibles amb l'objectiu de facilitar la distinció de cada tipus d'energia. L'única decisió amb la qual s'experimentarà, serà quin tipus de gràfic s'ajusta més a les necessitats de la informació mostrada. Com al cas anterior, es contemplen els gràfics de línies i els gràfics de barres. A la Figura 5.17 i 5.18 podem veure les gràfiques construïdes.

Per facilitat de lectura, s'ha optat per la gràfica de línies. Les agrupacions de la gràfica de barres poden confondre a l'usuari. Tot i que no és necessari, s'ha decidit incloure la correlació del tipus total per tal de facilitar la comparació d'aquesta amb les altres.

5.6 Generació de recomanacions

Per assolir l'objectiu principal del projecte, és necessari saber interpretar la informació extreta fins ara per aconseguir generar recomanacions de millora pels fàrmacs antiinflamatoris inhibidors de la COX-2. En aquest apartat, explicarem quina informació es recopila per la generació de recomanacions i de què 3 parts es compon aquest procés. Aquestes parts s'expliquen a continuació:

1. Confecció de la matriu: El primer pas consisteix en la confecció d'una

Figura 5.17: Gràfica de línies amb les correlacions entre els tipus energètics i l'energia total.

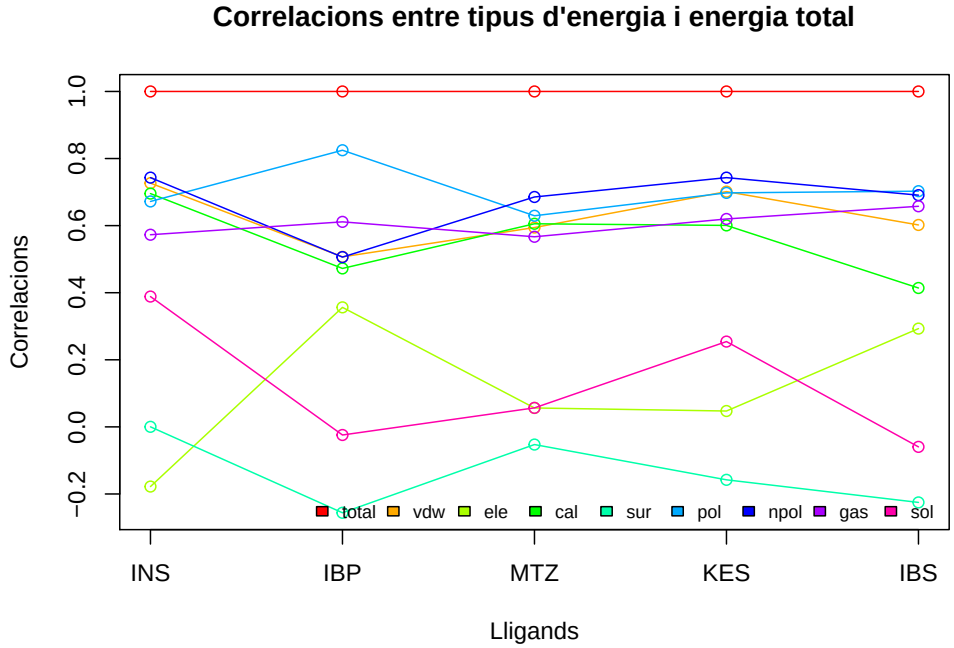
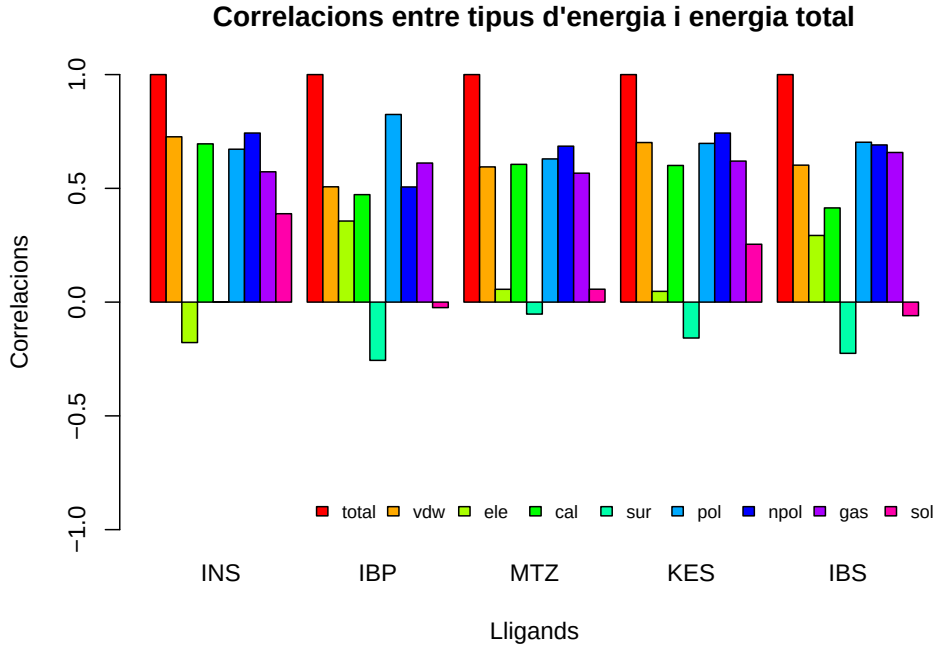


Figura 5.18: Gràfica de barres amb les correlacions entre els tipus energètics i l'energia total.



matriu de dades que contingui la informació necessària per fer possibles les següents fases.

2. Ordenament de la matriu: A continuació, s'ha de definir una puntuació capaç d'avaluar quin residu té més importància que la resta en funció de les dades de la matriu anterior.

3. Cerca dels àtoms més propers: Per últim, es realitza una cerca dels àtoms més propers a cada residu dins d'un model molecular.

El resultat d'aquest procés pretén ser una manera d'identificar quins residus són més importants a l'energia d'unió del fàrmac i quins àtoms del fàrmac interactuen més amb aquests residus. Als apartats següents s'explica més detalladament en què consisteix cada pas.

5.6.1 Confecció de la matriu de dades

La primera tasca d'aquest procés consisteix en la confecció de la matriu de dades, la qual ha de contenir informació suficient per a les fases posteriors. Com que l'objectiu és poder puntuar els residus de forma independent, les dades hauran de ser referents també als residus de forma individual. A més, per augmentar el nivell de detall, també s'ha considerat el fet d'analitzar cada residu per a cada tipus energètic disponible de forma separada, obviant el tipus corresponent al total de l'energia d'unió. Seguint aquestes premisses, s'han inclòs les següents columnes:

1. *Residu:* Correspon al nom del residu de la mostra. Per identificar-lo, utilitzarem el nom definit a l'apartat 5.2 del document. Aquests noms estan formats per 3 lletres majúscules, que determina el tipus del residu, seguides d'un número, que identifica el residu en qüestió.

2. *Tipus d'energia:* Es tracta d'una de les abreviacions dels tipus d'energia explicats a l'apartat 3.1 del document. Com ja s'ha especificat al principi de la secció 5.6.1, s'obviarà l'energia total d'unió.

3. *Mitjana de la contribució:* Fa referència a la mitjana de contribució del residu per tots els lligands analitzats i en el tipus d'energia especificat a la columna anterior.

4. *Desviació estàndard:* Es calcula computant la mitjana de la desviació de les dades energètiques del residu per cada lligand.

5. *Correlació de Pearson:* Correspon al coeficient de correlació entre la mitjana de la contribució total de cada lligand al llarg del temps i, la mitjana de la contribució del residu amb cada lligand i pel tipus energètic especificat. Es calcula utilitzant el coeficient de correlació de Pearson.

6. *Correlació de Spearman:* Respon a la mateixa descripció que la correlació de Pearson, excepte en el coeficient de correlació utilitzat. En aquest cas s'utilitza el coeficient de correlació de Spearman.

Tot i que les dues primeres columnes no tenen una especial complexitat, pot ser la resta sí que mereixen que les detallem amb cura.

La columna anomenada *mitjana de la contribució* correspon a la següent especificació: Donat un residu r, el tipus d'energia t i el conjunt total de lligands L. La contribució del residu r per a cada lligand l que pertany a L, és la mitjana de tots els valors de la columna del residu r a la matriu identificada pel lligand l i el tipus t. Llavors, el valor total és la mitjana de les contribucions del residu r per a cada lligand.

Per calcular la *desviació estàndard* es realitza el següent càlcul: Seguint el mateix vocabulari utilitzat a la definició de la *mitjana de la contribució*, la *desviació estàndard* del residu r per a cada lligand l que pertany a L, és la desviació dels valors de la columna del residu r a la matriu identificada pel lligand l i el tipus t. Llavors, el valor total és la mitjana de les desviacions del residu r per a cada lligand.

Per últim, expliquem el mètode de còmput de les correlacions. Aprofitem la mateixa explicació per descriure el mètode corresponent a la correlació de Pearson i la de Spearman. L'única diferència serà el coeficient de correlació utilitzat. El mètode és el següent: Seguint el mateix vocabulari utilitzat a la definició de la *mitjana de la contribució*, primer hem de considerar el vector de contribució total. Aquest vector correspon a la mitjana dels vectors de les contribucions totals de cada lligand. Una vegada dit això, la correlació del residu r per a cada lligand l que pertany a L, és el coeficient de

Figura 5.19: Fragment de la matriu de dades per a la generació de recomanacions.

Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.
ACE1	cal	-0.00048780487804877999949080	-0.12424478024012799992626	-0.14362164444
ACE1	ele	0.00040650406504065002668072	-0.10064751080988300346863	-0.10834785324
ACE1	gas	0.00040650406504065002668072	-0.10064751080988300346863	-0.10834785324
ACE1	pol	-0.00008130081300813050136297	-0.1578388467922970106 4607	-0.15515872402
ACE1	sol	-0.00048780487804878097527275	-0.05318905268488210297262	-0.07596721084
ACE554	cal	-0.00097560975609756097476355	-0.16815888360086500563639	-0.14884476071
ACE554	ele	0.00081300813008130005336144	-0.06531323914617409576167	-0.08645552797
ACE554	gas	0.00081300813008130005336144	-0.06531323914617409576167	-0.08645552797
ACE554	pol	-0.00016260162601626100271728	-0.17722413488049901242327	-0.17237594500
ACE554	sol	-0.00048780487804877999949080	0.1670939107930770051 1691	0.16593221869

correlació corresponent entre la columna del residu r a la matriu identificada pel lligand l i el vector de contribució total.

La raó per la qual calculem la correlació mitjançant dos coeficients diferents és per tal de detectar correlacions de diferents característiques. Com s'explica a l'apartat 3.3, el coeficient de Pearson només és capaç de trobar correlacions lineals. En canvi, amb el coeficient de Spearman, també es detecten correlacions monòtones, el que ens aporta més possibilitats.

El resultat d'aquests càlculs, ens dóna informació suficient per avaluar cada residu, a la Figura 5.19 podem veure un fragment de la matriu resultant. Respecte al tema de les correlacions, no s'aprecia una notable diferencia entre la correlació de Pearson i la de Spearman, pel que hem decidit utilitzar exclusivament la correlació de Pearson, tot i que seguirem incloent la de Spearman a la matriu.

5.6.2 Puntuació i ordenament dels residus

Una vegada obtinguda la matriu de dades, s'ha de definir la importància de cada mostra. El primer índex que es considera és la correlació, ja que l'objectiu és millorar l'energia d'unió del fàrmac millorant les contribucions energètiques d'alguns residus concrets. Una primera versió seria

Figura 5.20: Fragment de la matriu de dades ordenada segons el valor absolut de la correlació de Pearson.

	Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.
1	GLU334	gas	0.52406504065040604434244642	0.42125143290293398656843	0.39857819438
2	PRO266	ele	0.00268292682926828986167411	0.42006106101457302282043	0.42904284491
3	LYS790	ele	-0.17219512195121999820202063	0.4079018390938339844176l	0.40424994092
4	LYS790	gas	-0.17219512195121999820202063	0.4079018390938339844176l	0.40424994092
5	SER373	sol	-0.00073170731707317094791310	0.39977183677827798913285	0.36539024024
6	TYR355	ele	-0.03991869918699190306066171	0.39641539345176501907275	0.37420599010
7	GLU334	ele	0.52504065040650405027378156	0.39384611215039200260435	0.36802397173
8	ASP976	cal	-0.20853658536585401028773390	0.39255069326561597975456	0.38811871284
9	LYS703	sol	0.17804878048780500843228936	-0.39097408561494501055833	-0.38309591733
10	ARG90	gas	-6.49682926829267959334401894	0.36323570989205600234584	0.33337324091

ordenar la matriu segons el valor absolut de la correlació de cada mostra. Això ens permet detectar quins residus estan més relacionats amb l'augment o la disminució de l'energia d'unió del compost. A la Figura 5.20 podem veure els primers elements de la matriu ordenada pel valor absolut de la correlació de Pearson.

Si analitzem la matriu resultant, podem veure que ens aporta informació molt interessant respecte a les correlacions. Però ens adonem que estem obviant la informació relacionada a la contribució energètica, que és el valor que volem maximitzar. Per tant, una segona alternativa que es planteja és ordenar la matriu segons la mitjana de contribució de cada mostra. Com que en aquest cas, les energies més favorables són les més petites, ordenem la matriu segons aquest criteri. Un fragment del resultat es mostra a la Figura 5.21.

Amb aquesta nova versió, distingim molt fàcilment quins residus contribueixen a l'energia total d'unió i amb quin tipus d'energia ho fan. Però ara estem perdent la informació referent a la correlació. Volem plantejar una solució que aprofiti ambdues informacions, per això es proposa fer una funció que utilitzi les dues dades. Com que sabem que les correlacions sempre van de -1 a 1, però els valors de la contribució pertanyen a un ventall més ampli, normalitzarem els valors de la contribució perquè el valor màxim sigui 1. Com que només ens interessen les energies negatives, simplement dividirem la contribució energètica de cada mostra per la contribució energètica més

Figura 5.21: Fragment de la matriu de dades ordenada segons la mitjana de contribució.

	Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.
1	ARG90	gas	-6.4968292682926795933440189 4	0.36323570989205600234584	0.33337324091
2	ARG90	ele	-5.8707317073170699472939304 5	0.24910757155804200135840	0.22507496150
3	ALA497	gas	-2.2713008130081298041602622 1	0.11844625608375600189781	0.11073255753
4	ALA497	npol	-2.2036585365853698093019374 9	0.06984104260087639626242	0.06941156209
5	GLU494	cal	-2.1924390243902398545117193 9	0.08619109265208159398330	0.08466335658
6	GLU494	sol	-2.1909756097561001730866792 1	0.07285430485774420372724	0.07890826989
7	ALA497	vdw	-2.0480487804878100277505836 9	0.07072407274898849516287	0.06169093238
8	VAL319	npol	-1.8541463414634100637101710 2	0.15550459475008798770901	0.21401190918
9	VAL493	npol	-1.7458536585365900251076709 5	-0.11509652382407299653355	-0.11403679052
10	ARG483	gas	-1.7278861788617900785425263 2	0.08451796201878640502070	0.09048015549

Figura 5.22: Fragment de la matriu de dades ordenada segons la puntuació, calculada a partir de la mitjana de contribució i la correlació de Pearson.

	Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.	Puntuació
1	ARG90	gas	-6.4968292682926795933440189 4	0.36323570989205600234584	0.33337324091	1.86227768387
2	ARG90	ele	-5.8707317073170699472939304 5	0.24910757155804200135840	0.22507496150	1.49498156278
3	PRO266	ele	0.0026829268292682898167411	0.42006106101457302282043	0.42904284491	0.99676124156
4	LYS790	ele	-0.1721951219512199982020206 3	0.40790183909383398441761	0.40424994092	0.99481416363
5	LYS790	gas	-0.1721951219512199982020206 3	0.40790183909383398441761	0.40424994092	0.99481416363
6	LEU322	gas	-1.4675609756097600655522228 4	0.31812707252720701101012	0.32365858974	0.98108403074
7	ARG90	pol	-1.4124390243902400499109717 2	0.31787210921399799978815	0.27897712744	0.97199433971
8	ASP976	cal	-0.2085365853658540102877339 0	0.39255069326561597975456	0.38811871284	0.96396612140
9	LEU322	vdw	-1.4507317073170700183482040 3	0.30708709353156898869130	0.30209956461	0.95228607282
10	SER373	sol	-0.0007317073170731709479131 0	0.39977183677827798913285	0.36539024024	0.94912265957

petita de tota la matriu. D'aquesta manera obtenim un valor màxim d'u que denotarà la mostra amb mitjana de contribució més petita. La nova versió sumarà el valor de la correlació amb la mitjana de contribució normalitzada i ordenarà la matriu segons aquest valor. Per facilitar la comprensió de l'ordenament, afegirem la columna de *puntuació* a la matriu, que tindrà un valor màxim de 2 (màxima correlació i mínima contribució normalitzada son 1). La nova ordenació queda reflectida a la Figura 5.22.

La matriu resultant ens permet localitzar de forma rapida quins són els residus amb contribucions més favorables i que es correlacionen més amb l'energia total. Això facilita la llista de candidats en ordre d'importància per

tal de fer recomanacions. Així doncs, donem com a bona la funció de càlcul de la puntuació per realitzar l'ordenació.

5.6.3 Cerca dels àtoms més significatius

Amb la matriu generada, podem cercar quins àtoms del fàrmac són més significatius per cada mostra. L'àtom més significatiu per una mostra, serà el més proper al residu de la COX-2 que intervé. Aquesta anàlisi es pot realitzar fàcilment donat un model molecular tridimensional.

En el nostre cas, utilitzem una llibreria de R que es diu *Bio3D* [32]. Aquesta llibreria ens permet manipular un model molecular en format .pdb i realitzar operacions sobre els components d'aquest. A més, facilita el càlcul de matrius de distàncies, que ens serà d'utilitat durant el procés.

Abans d'explicar l'algoritme utilitzat, és necessari fer un petit incís per entendre'l. Un residu és una estructura atòmica, això vol dir que està format per un conjunt d'àtoms connectats entre ells. Però tots els àtoms no intervenen de la mateixa forma, de fet, hi ha un grup d'àtoms que no intervenen a la interacció. Aquests àtoms són els pertanyents a la cadena principal [33] ("Backbone chain" o "Main chain" en anglès). Per aquesta raó, no tindrem en compte aquests àtoms per calcular les distàncies.

El primer pas és seleccionar el lligand en el qual volem cercar els àtoms, aquest lligand ha de ser especificat abans de l'inici del procediment, així com el model molecular que es vol utilitzar. Una vegada fet això, repetim un algoritme per cada mostra de la matriu. A continuació expliquem que passos conformen el procés iteratiu:

1. *Identificador del residu:* En primer lloc, s'extreu l'identificador del residu del seu nom. Com ja s'ha explicat prèviament a l'apartat 5.2, basta amb eliminar els 3 primers caràcters del nom del residu per obtenir el seu identificador.

2. *Discriminació d'àtoms:* Una vegada identificat el residu, procedim a identificar els àtoms que el conformen. Bio3D ens facilita un mètode per, donat l'identificador de l'aminoàcid, extreure el conjunt d'àtoms

Figura 5.23: Fragment de la matriu de dades ordenada segons la puntuació, amb l'identificador i el tipus de l'àtom afectat per cada mostra.

	Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.	Puntuació	Id de l'àtom	Tipus de l'àtom
1	ARG90	gas	-6.4968292682926795933440189 4	0.3632357098920560023458 4	0.33337324091	1.86227768387	8901	HVT4
2	ARG90	ele	-5.8707317073170699472939304 5	0.2491075715580420013584 0	0.22507496150	1.49498156278	8901	HVT4
3	PRO266	ele	0.0026829268292682898616741 1	0.4200610610145730228204 3	0.42904284491	0.99676124156	8885	HVT2
4	LYS790	ele	-0.1721951219512199982020206 3	0.4079018390938339844176 1	0.40424994092	0.99481416363	8924	O2D
5	LYS790	gas	-0.1721951219512199982020206 3	0.4079018390938339844176 1	0.40424994092	0.99481416363	8924	O2D
6	LEU322	gas	-1.4675609756097600655522228 4	0.3181270725272070110101 2	0.32365858974	0.98108403074	8901	HVT4
7	ARG90	pol	-1.4124390243902400499109717 2	0.3178721092139979997881 5	0.27897712744	0.97199433971	8901	HVT4
8	ASP976	cal	-0.2085365853658540102877339 0	0.3925506932656159797545 6	0.38811871284	0.96396612140	8924	O2D
9	LEU322	vdw	-1.4507317073170700183482040 3	0.3070870935315689886913 0	0.30209956461	0.95228607282	8901	HVT4
10	SER373	sol	-0.0007317073170731709479131 0	0.3997718367782779891328 5	0.36539024024	0.94912265957	8884	HVC2

que el conformen. Com ja hem explicat anteriorment, s'ha d'excloure alguns d'aquests, concretament els que conformen la cadena principal. La funcionalitat de selecció d'àtoms d'un aminoàcid, ens permet obtenir exclusivament els que pertanyen a aquesta cadena. Amb una simple diferència dels dos conjunts, obtenim els àtoms que componen el residu i que no pertanyen a la seva cadena principal.

3. *Càlcul de distàncies:* Per fer el càlcul de les distàncies, fem servir les coordenades tridimensionals que el model molecular ens facilita. Amb aquestes posicions, calculem la matriu de distàncies entre els àtoms de l'aminoàcid i els àtoms del fàrmac. Com que no diferenciem segons l'àtom del residu, la distancia entre aquest i cada àtom del fàrmac serà la mitjana de la distància de tots els components del residu amb cada àtom del compost.

4. *Identificació de l'àtom:* Per finalitzar, busquem quin és l'àtom que correspon a la distància més petita de les prèviament calculades. Aquest àtom serà el que s'ha de potenciar per tal de millorar l'energia d'unió de la mostra.

L'identificador de l'àtom més proper i el seu tipus seran inclosos a la matriu de dades. D'aquesta manera, en cada fila de la matriu podem observar una mostra amb totes les seves dades i quin és l'àtom del fàrmac més afectat. A la Figura 5.23 es mostra un fragment de la matriu resultant.

5.7 Desenvolupament d'una aplicació web

Com a última tasca, s'ha desenvolupat una aplicació web per facilitar l'ús del software desenvolupat. D'aquesta manera, s'ofereix una interfície gràfica que disminueix la corba d'aprenentatge que requeriria utilitzar una eina basada en scripts de R. Com ja vam especificar a l'apartat 5.1, on parlàvem de la selecció d'eines de desenvolupament, l'aplicació web serà implementada amb el framework Shiny. Per començar, hem definit quines funcionalitats ha d'oferir l'aplicació:

- *Visualització de les gràfiques de perfils globals per lligand:* En primer lloc, s'han de mostrar les 3 gràfiques de perfils globals per lligand que s'han dissenyat a l'apartat 5.5.1. Per visualitzar aquestes gràfiques, s'ha d'oferir la possibilitat de seleccionar quins fàrmacs es volen incloure.
- *Visualització de les gràfiques de perfils energètics per residu:* La segona funcionalitat és mostrar les 8 gràfiques de perfils energètics per residu que s'han dissenyat a l'apartat 5.5.2. Per visualitzar aquestes gràfiques, s'ha d'oferir la possibilitat de seleccionar quins fàrmacs es volen incloure, així com quin tipus d'energia es vol mostrar.
- *Visualització de la matriu de recomanacions:* Per últim, l'aplicació ha de ser capaç de mostrar la matriu de recomanacions explicada a l'apartat 5.6. Aquesta matriu ha de permetre dues versions, una primera sense indicar l'àtom més proper i el seu tipus, i una segona que mostri aquesta informació. Com que el sistema permet a l'usuari que seleccioni el model molecular sobre el qual vol que es realitzi la cerca d'àtoms més propers i especificar l'identificador del lligand en el model, es mostrarà una versió o un altre en funció de si es disposa de la informació necessària (model molecular i identificador del lligand al model). Per altra banda, també s'ha d'oferir la possibilitat de seleccionar quins fàrmacs dels disponibles es volen incloure al càlcul de la matriu.

Una vegada estudiades les necessitats del sistema, s'ha de decidir quina serà l'estructura de l'aplicació. Com que es vol utilitzar alguna estruc-

tura estàndard de Shiny, s'ha decidit la utilització de la distribució "pageWithSidebar". Aquesta distribució divideix l'aplicació en 3 parts:

- *Títol:* Es mostra a la part superior de l'aplicació i conté el títol d'aquesta.
- *Panel d'entrada:* Ocupa la part esquerra del sistema i permet la introducció de dades per l'usuari.
- *Panel de sortida:* Ocupa la part dreta de l'eina i és on es mostren les diferents visualitzacions desitjades.

Aquesta estructura encaixa perfectament amb els requisits de l'aplicació. Tot i que encara falta definir el contingut de cada part. A causa de la simplicitat de l'apartat del títol, obviem la seva descripció. A continuació expliquem els panels d'entrada i sortida.

5.7.1 Panel d'entrada

En aquesta part de l'aplicació s'inclouen tots els controls dels quals disposa l'usuari per influir en la resposta del sistema. Si estudiem els requisits definits a l'apartat anterior, detectem les següents necessitats:

- *Seleccionar els compostos:* L'usuari ha de poder seleccionar el subconjunt de compostos que intervenen en les 3 visualitzacions generades. El conjunt de compostos disponibles vindrà definit per les dades de les quals disposi el sistema.
- *Seleccionar el tipus d'energia:* L'usuari ha de poder seleccionar el tipus d'energia utilitzat per la generació de les dues gràfiques. En tot moment haurà d'haver-hi un tipus energètic seleccionat i només un. A més, aquest haurà de ser un dels establerts en el projecte (en total son 9).
- *Introduir l'identificador de l'àtom al model molecular:* La persona que utilitzi el sistema ha de poder introduir l'identificador numèric de l'àtom al model molecular. Aquest identificador s'utilitzarà al càlcul dels àtoms més propers que s'inclouen a la matriu de recomanacions. Com en el cas anterior, aquest camp mai podrà ser buit.

- *Aportar un model molecular:* Per últim, l'usuari ha de poder pujar un model molecular tridimensional en forma de fitxer amb extensió .pdb. La pujada d'aquest fitxer és opcional, ja que és el que determina la versió de la matriu de recomanacions que es mostra. Aquest model serà en el que es basi la cerca d'àtoms més propers a cada residu de la proteïna.

Una vegada definides les necessitats del panel d'entrada, se seleccionen els components que millor les resolen:

- *Seleccionar els compostos:* Per seleccionar els compostos s'ha decidit utilitzar un "selectInput". En aquest tipus de component, s'especifica un conjunt d'opcions i es pot configurar de manera que permeti la selecció múltiple.
- *Seleccionar el tipus d'energia:* Per seleccionar el tipus d'energia s'utilitza el mateix component que en el cas anterior. L'única diferència és que en aquesta ocasió no es permetrà la selecció múltiple.
- *Introduir l'identificador de l'àtom al model molecular:* Com que l'identificador és un número, s'utilitzarà un component anomenat "numericInput". Aquest només permet l'entrada de números, evitant possibles errors per entrades incorrectes. A més, permet especificar un valor per defecte, on es pot incloure l'identificador utilitzat per defecte en aquest tipus de models.
- *Aportar un model molecular:* Finalment, per la pujada de fitxers existeix un component anomenat "fileInput". Aquest permet la selecció d'un fitxer mitjançant una vista de navegació pel sistema de fitxers local i el carrega automàticament.

Amb els components més adequats seleccionats, ja podem implementar la part d'entrada de dades. Podem veure una captura de l'aspecte de l'aplicació a les Figures 10.12, 10.13, 10.14 i 10.15 de l'annex, on aquesta part se situa a l'esquerra del disseny.

5.7.2 Panel de sortida

L'altra part important de l'aplicació és el panel de sortida, en aquest s'inclouen totes les modalitats de visualització que ofereix el sistema. Si estudiem els requisits definits a l'apartat 5.7, detectem les següents necessitats:

- *Visualització de les gràfiques de perfils globals per lligand:* L'aplicació ha de poder visualitzar simultàniament els 3 gràfics de barres que corresponen a les gràfiques de perfils globals per lligand.
- *Visualització de les gràfiques de perfils energètics per residu:* El sistema ha de ser capaç de mostrar un gràfic de línies que representa la gràfica de perfils energètics per residu per un tipus d'energia seleccionat.
- *Visualització de la matriu de recomanacions:* Per últim, l'eina ha de permetre la visualització d'una matriu de dades d'una mida considerable (d'unes 6.000 files).

Una vegada definides les necessitats del panel de sortida, se seleccionen els components que millor les solucionen:

- *Visualització de les gràfiques de perfils globals per lligand:* Per a la visualització de gràfiques, Shiny proposa l'ús dels "plotOutput". Com que necessitem mostrar 3 gràfiques, però no volem perdre qualitat reduint la mida, es mostraran una sota l'altre.
- *Visualització de les gràfiques de perfils energètics per residu:* Com que es tracta d'una gràfica, utilitzarem el mateix component que hem comentat a l'apartat anterior.
- *Visualització de la matriu de recomanacions:* Per visualitzar taules de dades, hem decidit utilitzar un component anomenat "dataTableOutput". Aquest, a banda de permetre mostrar una taula de dades, ofereix funcionalitats extres com un cercador d'elements, l'ordenació per columnes o la divisió de la taula en pagines.

En aquest panel, es planteja el problema afegit del fet que volem mostrar les 3 parts a la vegada, però no volem limitar la mida de cadascuna

per no perdre qualitat de visualització. Per resoldre aquest requisit, definirem un component anomenat "tabsetPanel" que permet afegir diverses vistes a la vegada i ofereix una barra de navegació per canviar de vista de forma àgil. A les Figures 10.12, 10.13, 10.14 i 10.15 de l'annex, es mostra l'aspecte de l'aplicació, on el panel de sortida es mostra a la dreta del disseny.

6 Resultats

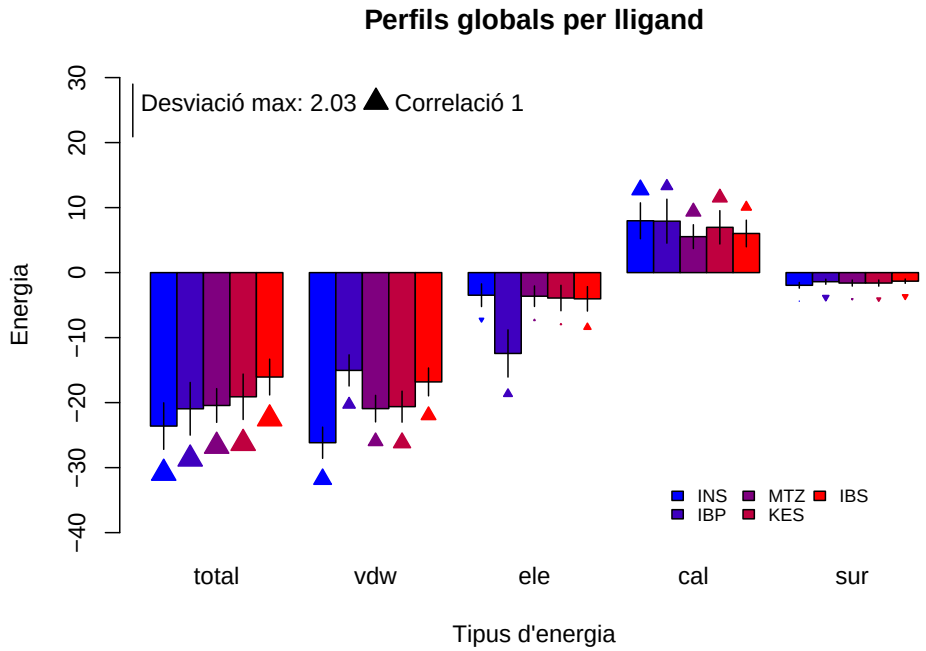
El resultat del projecte és un software que, partint d'un conjunt de dades que descriuen la interacció de diferents fàrmacs sobre la COX-2, és capaç de generar una sèrie de gràfiques que ajuden a la comprensió de les dades i d'un llistat de possibles punt de millora del fàrmac. Les dades han d'estar en un format concret, descrit a l'apartat 5.2. A continuació, es fa una especificació dels resultats concrets.

6.1 Perfils globals per lligand

La primera gràfica pretén donar una visió global de com interactua cada fàrmac amb la proteïna. Aquesta informació es quantifica segons l'energia d'unió que té el fàrmac amb la COX-2. També es pot observar la descomposició d'aquesta energia segons els diferents tipus d'energia, els quals estan explicats a l'apartat 3.1. Com que contemplen 3 descomposicions diferents de l'energia total, es generen 3 gràfiques on cadascuna mostra una d'aquestes. Podem veure l'exemple per als 5 fàrmacs utilitzats i la primera descomposició de l'energia a la Figura 6.1.

A la interpretació de la gràfica, hem de tenir en compte que, com més petita és l'energia d'unió del fàrmac amb la proteïna, més afavorida es veu la interacció entre aquests. Dit això, a la gràfica podem veure com influeix cada tipus d'energia (Eix horitzontal) a l'energia d'unió (Eix vertical). Tot això, replicat per cada lligand diferent, els quals s'identifiquen per colors. El color de cada lligand, el qual s'especifica a la llegenda de la part inferior,

Figura 6.1: Gràfica de perfils energètics globals per lligand referent a la primera descomposició de l'energia total.



és un gradient de blau a vermell, on el blau més intens denota una energia d'unió més favorable (més petita) i el vermell és l'energia menys favorable (més gran).

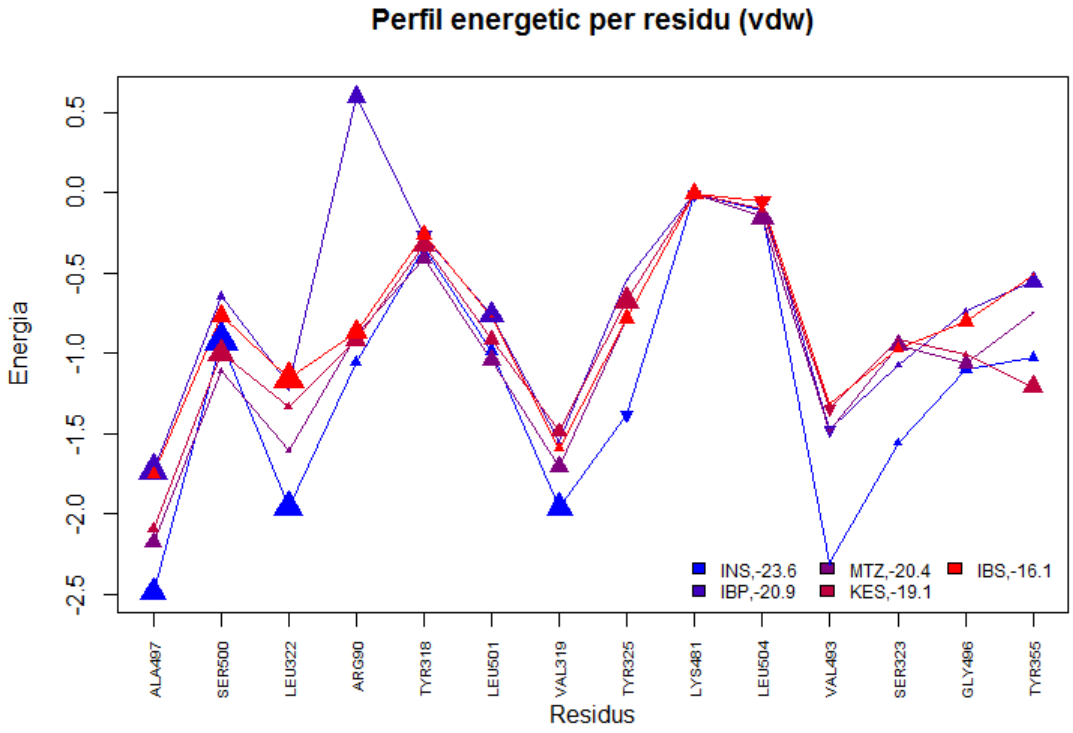
A banda de l'energia, també es mostra la desviació estàndard i la correlació de les dades energètiques. La desviació estàndard es representa amb una línia a l'extrem de cada barra. La mida de la línia, creix amb la desviació. A la llegenda superior podem veure quina és la desviació màxima que conté el gràfic i quina és la línia que la representa. Per mostrar la correlació de les dades de cada tipus respecte l'energia total es representa mitjançant un triangle a l'extrem de cada barra. La mida del triangle creix proporcionalment amb el valor absolut de la correlació. La posició del triangle denota el signe de la correlació, el fet que apunti cap a dalt representa una correlació positiva, mentre que si apunta cap a baix, la correlació es negativa. A la llegenda superior s'inclou la mida del triangle que representa el valor absolut de la correlació màxima o mínima.

Per últim, comentem els resultats obtinguts. Podem detectar que el lligand anomenat INS és el que millor energia d'unió genera. En contraposició al IBS, que és el que menys interacciona amb la proteïna. Per altra banda, podem veure que el tipus energètic més important és l'energia de Van der Waals, explicada a l'apartat 3.1. També és interessant veure com tots els fàrmacs presenten una descomposició energètica similar, excepte el IBP, que potencia més les unions d'energia electrostàtica. Als annexos 10.1, 10.2 i 10.3 podem veure la gràfica mostrada i les dues gràfiques que corresponen a les descomposicions energètiques segons polaritat i solubilitat de l'energia.

6.2 Perfils energètics per residu

La segona gràfica, expressa els residus més significatius per l'energia d'unió total de cada fàrmac per un tipus d'energia concret. Aquesta informació s'aproxima més a l'objectiu principal del projecte, ja que identifica els residus a través dels quals es poden introduir millores. La significança d'un residu ve donada per la seva mitjana de contribució energètica i per la correlació de la seva energia amb l'energia total. Si obviem l'energia interna (int), que sempre és nul·la, i la total, disposem de 8 tipus d'energies per

Figura 6.2: Gràfica de perfils energètics per residu referent a l'energia de Van der Waals.



analitzar i, per tant, generem 8 gràfiques diferents. La selecció d'aminoàcids que s'inclouen a la gràfica és la unió dels 10 aminoàcids amb una millor contribució i els 10 amb major correlació en valor absolut amb l'energia total. És per això, que depèn del tipus d'energia es mostren més o menys residus, ja que depèn de la quantitat d'aquests que es trobin entre els més contributius i els més correlacionats. A la Figura 6.2 podem veure la gràfica del perfil energètic per residu corresponent a l'energia de Van der Waals, ja que és l'energia més rellevant de la primera descomposició energètica segons la gràfica de perfils globals.

A la gràfica podem veure com contribueix cada residu de la proteïna (Eix horitzontal) a l'energia d'unió (Eix vertical) del tipus especificat. Tot

això, replicat per cada lligand diferent, els quals s'identifiquen per colors. L'assignació de colors és la mateixa que a les gràfiques de perfils globals per lligand. A més de la contribució energètica, també s'inclou la correlació de l'energia del residu respecte a l'energia total d'unió. Per representar la correlació, s'utilitza el mateix mètode basat en la mida i l'orientació dels triangles situats en cada intersecció entre l'eix d'un residu i d'un valor energètic. Per facilitar una millor apreciació de la influència de cada aminoàcid, s'ha inclòs el valor de l'energia total d'unió a la llegenda inferior.

Per finalitzar, valorem el resultat de la gràfica generada. Com a la gràfica anterior, veiem que el lligand INS és el que un millor model d'unió presenta, en contra del lligand IBS que presenta una interacció menys favorable. Fixant-nos en els residus, s'aprecia que els residus anomenats ALA497, VAL319 i VAL493 contribueixen de manera important en tots els lligands. Els aminoàcids identificats com TYR318, LYS481 i LEU504 presenten el cas contrari, aquests són poc significatius en la contribució total. Crida l'atenció el cas del ARG90, que presenta resultats més que acceptables per tots els lligands excepte pel IBP. Aquest fet és a causa de la diferència significativa que hem detectat a la gràfica de perfil global, on l'IBP mostra una energia d'unió menor respecte a la resta per l'energia de Van der Waals. Tot i que parlem d'energies menys significatives, hem de tenir en compte que totes les energies estan per sota de 0, i per tant son energies favorables a la interacció.

Si passem a valorar les correlacions visualitzades, podem distingir els residus amb correlacions similars entre lligands i altres que mostren diferencies molt grans d'un lligand a un altre. Exemple del primer cas seria el TYR318 i el VAL493. Per l'altra banda, tindríem el ALA497, el SER500 i el LEU322. Els del primer grup solen formar part d'un conjunt base que comparteixen la majoria dels fàrmacs del mateix tipus, a diferència dels del segon grup, que són els que acaben de potenciar l'efecte de cada fàrmac.

Als annexos entre el 10.4 i el 10.11 (ambdós inclosos) es poden veure les 8 gràfiques resultants corresponents als 8 tipus d'energia.

6.3 Matriu de recomanacions

Per últim, tenim la matriu de recomanacions. Es tracta d'una matriu de més de 6.000 files, on cada fila representa una mostra. Cada mostra descriu l'activitat d'un residu per un tipus d'energia i per a tots els lligands inclosos a l'estudi. En total, la matriu té 9 columnes, que representen:

1. L'identificador del residu.
2. El tipus d'energia.
3. La mitjana de contribució.
4. La desviació de les contribucions.
5. La correlació de Pearson entre les contribucions i l'energia total d'unió.
6. La correlació de Spearman entre les contribucions i l'energia total d'unió.
7. La puntuació de rellevància de la mostra.
8. L'identificador de l'àtom del fàrmac més proper.
9. El tipus de l'àtom del fàrmac més proper.

A l'apartat 5.6 podem veure l'explicació de com s'obté cada dada.

Aquesta matriu ha de ser interpretada com una llista de punts de millora dels fàrmacs antiinflamatoris inhibidors de la COX-2. Per facilitar la comprensió del que això comporta, farem un exemple amb un fragment de la matriu, que podem veure a la Figura 6.3. En aquestes primeres 10 mostres, podem veure que hi ha alguns residus que apareixen més d'un cop. Concretament el ARG90 apareix en 3 tipus energètics diferents en les 7 primeres posicions, això denota una gran importància d'aquest aminoàcid a l'eficàcia dels 5 fàrmacs analitzats. Comprenent la fórmula utilitzada per calcular la puntuació, explicada a l'apartat 5.9, notem que el factor determinant és la mitjana de contribució. El valor d'aquesta és la més alta que podem detectar a la matriu, tot i que també influeix un bon coeficient de correlació. Si veiem la gràfica dels perfils energètics per residu (Figura 10.4), veiem que

Figura 6.3: Fragment de la matriu de dades ordenada segons la puntuació amb informació dels àtoms més propers al residu.

	Residu	Tipus.d.energia	Mitjana.de.la.contribució	Correlació.Pearson.	Correlació.Spearman.	Puntuació	Id de l'àtom	Tipus de l'àtom
1	ARG90	gas	-6.4968292682926759334401894	0.363235709892056002345 84	0.33337324091	1.86227768387	8901	HVT4
2	ARG90	ele	-5.8707317073170699472939304 5	0.249107571558042001358 40	0.22507496150	1.49498156278	8901	HVT4
3	PRO266	ele	0.0026829268292682898616741 1	0.420061061014573022820 43	0.42904284491	0.99676124156	8885	HVT2
4	LYS790	ele	-0.1721951219512199982020206 3	0.407901839093833984417 61	0.40424994092	0.99481416363	8924	O2D
5	LYS790	gas	-0.1721951219512199982020206 3	0.407901839093833984417 61	0.40424994092	0.99481416363	8924	O2D
6	LEU322	gas	-1.4675609756097600655522228 4	0.318127072527207011010 12	0.32365858974	0.98108403074	8901	HVT4
7	ARG90	pol	-1.4124390243902400499109717 2	0.317872109213997999788 15	0.27897712744	0.97199433971	8901	HVT4
8	ASP976	cal	-0.2085365853658540102877339 0	0.392550693265615979754 56	0.38811871284	0.96396612140	8924	O2D
9	LEU322	vdw	-1.4507317073170700183482040 3	0.307087093531568988691 30	0.30209956461	0.95228607282	8901	HVT4
10	SER373	sol	-0.0007317073170731709479131 0	0.399771836778277989132 85	0.36539024024	0.94912265957	8884	HVC2

el residu amb una contribució més important de l'energia electrostàtica és precisament aquest, en gran part determinat per la seva contribució amb l'IBP. Per aquesta raó, l'aminoàcid destaca en les 3 energies que depenen de l'energia electrostàtica. Però el que realment podem entendre com una recomanació és la localització de l'àtom del fàrmac més proper. Aquesta mostra ens determina que substituint l'àtom amb identificador 8901 per un altre que afavoreixi l'energia electrostàtica, s'augmentaria l'eficàcia del fàrmac.

7 Conclusions i treball futur

En aquest projecte s'ha desenvolupat un software que, partint d'un conjunt de dades que descriuen la interacció de diferents fàrmacs sobre la COX-2, és capaç de generar una sèrie de gràfiques que ajuden a la comprensió de les dades i d'un llistat de possibles punts de millora del fàrmac. Aquest pretén ser una eina de suport per a la fase de disseny molecular d'un nou fàrmac antiinflamatori inhibidor de la COX-2. En primer lloc, comentarem quines conclusions globals s'han extret de la realització del projecte i s'identificaran possibles tasques futures per millorar el software.

7.1 Conclusions

Per extreure unes conclusions objectives del desenvolupament del projecte, analitzarem si els objectius que es plantejaven (apartat 2.2) han sigut assolits, o no. En cas afirmatiu, s'explicarà com, i en cas negatiu, es raonarà el perquè. Els objectius són els següents:

- **Extreure informació:** El primer objectiu era ser capaços d'extreure informació útil a partir de l'anàlisi de les dades de les quals disposem. Aquest objectiu ha sigut assolit, ja que, com s'explica a l'apartat 5.4, s'ha aconseguit construir una sèrie de matrius de dades que aporten informació útil per a la consecució de la resta d'objectius. A banda d'aquest apartat, podem considerar l'apartat de construcció de la matriu de recomanacions (apartat 5.6) com altre bon exemple d'extracció d'informació d'utilitat.

- **Representar gràficament la informació:** El següent objectiu, era representar amb gràfics de diferents tipus la informació extreta. Aquest també s'ha completat amb èxit, ja que es generen les gràfiques de perfils globals per lligand (apartat 5.5.1) i les de perfils energètics per residu (apartat 5.5.2). Les gràfiques han anat evolucionant durant el desenvolupament per tal d'adaptar-les al seu objectiu concret i millorar la seva expressivitat. Finalment, s'ha generat una versió d'ambdues gràfiques, eficaç i de fàcil comprensió per l'usuari.

- **Fer suggeriments per la millora d'inhibidors de la COX-2:** A continuació, es plantejava la generació de suggeriments de quins residus és recomanable potenciar o debilitar per afavorir la unió total del fàrmac amb la proteïna. L'algoritme de generació de recomanacions ha sigut implementat de forma satisfactòria, tal com s'explica a l'apartat 5.6. A més, els resultats obtinguts, avaluats a l'apartat 6.3, també són de bona qualitat.

- **Oferir una interacció adequada:** I per finalitzar, es pretenia desenvolupar algun tipus d'interfície simple i funcional per facilitar l'ús de l'eina als usuaris. Amb el desenvolupament de l'aplicació web explicada a l'apartat 5.7, s'ha aconseguit completar aquest objectiu. El sistema permet accedir a totes les funcionalitats implementades al projecte d'una forma simple i funcional. A més, els temps de resposta finals són molt acceptables per a la quantitat de dades analitzades. La fase d'optimització dels processos s'ha considerat de màxima prioritat per oferir una interacció acceptable a l'usuari sense temps d'espera molt llargs.

Fent referencia a la planificació temporal, és interessant comentar algunes contradiccions que presenta enfront de la metodologia àgil utilitzada (apartat 8.1). Les metodologies àgils, permeten acotar l'abast d'un projecte per temps o per funcionalitats, però no per ambdues alhora. Donat que aquest projecte té una data límit i requereix una definició prèvia de l'abast, pot ser, la metodologia seleccionada no és la millor pel tipus de projecte. Dit això, val la pena destacar que aquest problema no ha afectat l'assoliment d'objectius ni a derivat en cap endarreriment temporal de la finalització del treball.

7.2 Treball futur

En aquesta secció, es vol definir un conjunt de possibles tasques futures que, poden ajudar a la millora i l'evolució de l'eina resultant d'aquest projecte. És interessant identificar aquest tipus de tasques per facilitar la realització d'una segona fase de desenvolupament o impulsar a la seva realització. A continuació s'enumeren els punts de millora:

- *Període de proves:* Una possible tasca a realitzar seria la realització d'un període de proves amb la participació dels usuaris finals. Analitzar com ells treballen amb l'eina pot identificar punts de millora de l'aplicació. A més, els mateixos usuaris poden transmetre idees d'utilitat en les quals l'equip no hagi pensat prèviament.

- *Calibrar puntuació de la matriu de correlacions:* La funció que calcula la puntuació actualment ho fa a partir de la mitjana de contribució i la correlació de les dades, donant el mateix pes als dos factors. Possiblement, aquesta funció podria donar millors resultats si es distribuïssin els pesos d'altra manera, o es canviessin els factors utilitzats. El fet de no tenir una manera objectiva d'avaluar la qualitat dels resultats, donat que no existeix una resposta correcta predefinida, dificulta la tasca de provar diferents mètriques per quantificar el seu grau de qualitat. És possible que la realització d'aquesta tasca requereixi l'ajut dels usuaris especialistes en la matèria, fent referencia al primer punt d'aquesta llista.

- *Implementar algoritmes de predicció més complexos:* Donat la limitació temporal del projecte, no ha sigut possible la implementació d'alguna tècnica avançada de predicció de conseqüències de canvis específics sobre el fàrmac. Aquestes tècniques, de les quals algunes han sigut comentades a l'estat de l'art del projecte (apartat 4.2), potser serien capaces de generar altres recomanacions o avaluar prèviament les ja generades.

- *Millorar la interfície gràfica de l'aplicació web:* Pel mateix motiu que a l'apartat anterior, no s'ha pogut dedicar molts esforços a la part de disseny de la interfície web. Tot i que l'objectiu de l'aplicació era facilitar l'ús de l'eina d'una forma simple i funcional, pot ser un millor

disseny amb la col·laboració d'un especialista en aquest àmbit podria millorar l'experiència de l'usuari.

8 Gestió del projecte

8.1 Metodologies de treball

A l'hora de plantejar un projecte nou, és important pensar que volem fer (objectius), però també com ho volem fer. Per això, una decisió important és la identificació de quines metodologies de treball volem utilitzar. En aquest apartat, parlarem sobre aquest tema. La metodologia que s'utilitzarà en aquest projecte és una metodologia àgil [34] anomenada Scrum. En aquest apartat s'ha inclòs també el mètode que utilitzarem de control de versions. A continuació expliquem tots aquests conceptes i en què ens ajuden.

8.1.1 Metodologies àgils

En 2001, 17 especialistes en noves metodologies i crítics amb la forma tradicional de treballar en el desenvolupament del software es van reunir. La reunió va ser impulsada per Kent Beck, enginyer de software nord-americà i un dels principals creadors del "extrem programming" [35] i el TDD[36]. Els integrants d'aquest grup de professionals, amb una dilatada experiència en el món del desenvolupament del software, portaven aproximadament una dècada destacant en aquesta indústria gràcies a l'ús de tècniques innovadores. Els unia el fet que no creien adient el clàssic model en cascada on primer s'analitza, després es dissenya, posteriorment s'implementa i per últim es testeja el producte. El resultat de la reunió va ser l'anomenat manifest àgil [37], el qual es compon per 4 punts que citem a continuació:

> Estem descobrint millors maneres de desenvolupar software tant per la nostra pròpia experiència com ajudant a tercers. A través d'aquesta experiència hem après a valorar:
>
> - **Individus i interaccions** sobre processos i eines.
> - **Software que funciona** sobre documentació exhaustiva.
> - **Col·laboració amb el client** sobre negociació de contractes.
> - **Respondre davant del canvi** sobre seguiment d'un pla.
>
> Això és, tot i que els elements a la dreta tenen valor, nosaltres valorem per sobre d'ells els de l'esquerra.
>
> *Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, Dave Thomas.*

A partir d'aquest manifest, es van definir els 12 principis que s'han de seguir per desenvolupar software de forma àgil:

1. La nostra major prioritat és satisfer al client mitjançant l'entrega primerenca i continua de software amb valor.
2. Acceptem que els requisits canvien, fins i tot en etapes tardanes del desenvolupament. Els processos àgils aprofiten el canvi per proporcionar avantatge competitiu al client.
3. Entreguem software funcional freqüentment, entre dues setmanes i dos mesos, amb preferència pel període de temps més curt possible.
4. Els responsables de negoci i els desenvolupadors treballen junts de forma quotidiana durant tot el projecte.
5. Els projectes es desenvolupen al voltant d'individus motivats. Cal donar-los l'entorn i el suport que necessiten, i confiar-los l'execució de la feina.

6. El mètode més eficient i efectiu de comunicar informació a l'equip de desenvolupament i entre els seus membres és la conversació cara a cara.

7. El software funcionant és la mesura principal de progrés.

8. Els processos àgils promouen el desenvolupament sostenible. Els promotors, desenvolupadors i usuaris devem ser capaços de mantenir un ritme constant de forma indefinida.

9. L'atenció contínua és l'excel·lència tècnica i el bon disseny millora l'agilitat.

10. La simplicitat, o l'art de maximitzar la quantitat de feina no realitzada, és essencial.

11. Les millors arquitectures, requisits i dissenys emergeixen d'equips autoorganitzats.

12. A intervals regulars l'equip reflexiona sobre com ser més efectiu per a continuació ajustar i perfeccionar el seu comportament en conseqüència.

Una vegada definides les principals característiques de les metodologies àgils, podem explicar amb més detall la metodologia Scrum [38], que és la que nosaltres hem seleccionat.

8.1.2 Scrum

Scrum va ser identificat i definit per Ikujiro Nonaka i Hirotaka Takeuchi a principis dels anys 80. L'idea va sorgir en analitzar la metodologia emprada per les principals empreses tecnològiques al seu procés de desenvolupament de productes. El nom de Scrum, melé en angles (Figura 8.1), prové de la comparació [39] que van fer d'un mètode de desenvolupament semblant al d'una melé de rugbi enfront d'un semblant a una cursa de relleus:

> L'enfocament de 'cursa de relleus' en el desenvolupament de productes [...] pot entrar en conflicte amb els objectius de màxima velocitat i flexibilitat. En el seu lloc, un enfocament holístic o

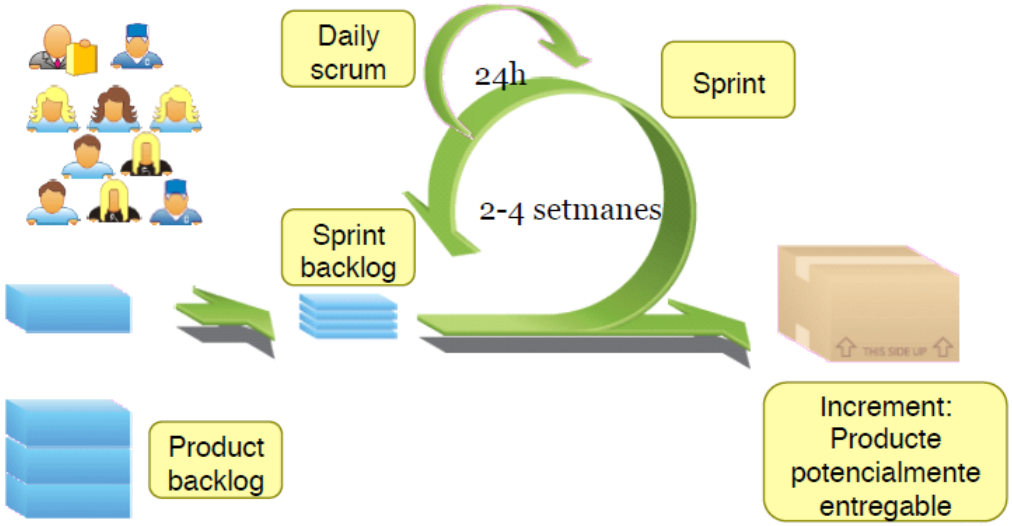
Figura 8.1: Dos equips participant en una melé (Scrum en anglès) durant un partit de rugbi.

> estil 'rugbi' - on un equip intenta anar a la distància com una unitat, passant la pilota cap endavant i cap enrere - poden servir millor als actuals requisits competitius.
>
> *Ikujiro Nonaka i Hirotaka Takeuchi*

Una de les característiques més importants de Scrum és la divisió del desenvolupament en esprints. Un esprint és un període de temps d'entre 2 i 4 setmanes de duració que completa un cicle al procés de Scrum, com podem veure a la Figura 8.2. Per entendre millor aquesta metodologia, definim els 3 rols, 4 cerimònies i 3 artefactes que la componen.

1. **Rols:** Les responsabilitats d'un projecte Scrum es divideixen en 3 rols:
   (a) *Product Owner (PO, client):* És el responsable de maximitzar el valor del producte representant els interessos dels usuaris finals. S'encarrega de definir les funcionalitats del producte final,

Figura 8.2: Esquema visual del procés de desenvolupament de Scrum.



les dates i continguts de les versions i la prioritat de cada funcionalitat. Això ho fa mitjançant el Product Backlog (explicat a la secció d'artefactes), expressant clarament el contingut per tal que el DT ho entengui sense problemes.

(b) *Development Team (DT, equip de desenvolupament):* És l'equip de desenvolupadors que s'encarrega de dur a terme el projecte. Es tracta d'un equip (d'entre 3 i 9 persones) autoorganitzat, sense diferències jeràrquiques i sense sub-equips. Altre aspecte important és que l'equip no es divideix segons el perfil tècnic de cadascú, tots responen com a equip de la seva feina.

(c) *Scrum Master (SM):* És el responsable del bon funcionament de tot el procés. S'encarrega de detectar i corregir defectes en la metodologia emprada, per aconseguir una millora continua. Altra funció és ajudar a tots els actors del procés a entendre i executar correctament les seves funcions en aquest.

2. **Cerimònies:** Scrum requereix la celebració de 4 tipus de reunions per assegurar un funcionament correcte de la metodologia. En totes elles participen els 3 rols definits anteriorment:

(a) *Sprint planning:* Consisteix a definir la feina a realitzar durant el següent esprint. El PO explica les funcionalitats no implementades més prioritàries al DT. El DT estima la feina que porta cada una de les funcionalitats per ajudar al PO a fer-se una idea de les dimensions de cadascuna. Una vegada coneguda la prioritat i l'esforç requerit de cada funcionalitat, el PO decideix quines s'han de desenvolupar en el següent esprint.

(b) *Daily scrum:* Es tracta d'una reunió diària on cadascú dels membres de l'equip exposa la seva resposta a 3 preguntes: "Què has fet des de l'última reunió Daily Scrum?", "Què planeges fer fins al següent Daily Scrum?" i "Quins problemes hi ha que puguin perjudicar l'Scrum actual o al projecte?". D'aquesta manera, el DT sincronitza activitats i planifica el dia.

(c) *Sprint review:* Se celebra a la finalització de cada esprint i consisteix que el PO identifiqui quines tasques han sigut realitzades i quines no. A més, l'Scrum master discuteix què ha anat bé, quins problemes van sorgir i com es van solucionar. Sol celebrar-se immediatament abans que l'Sprint planning.

(d) *Sprint retrospective:* L'equip Scrum fa introspecció i planifica la millora del procés. Amb aquesta finalitat, s'inspecciona com va ser l'esprint pel que fa a les persones, relacions, processos i eines. S'identifiquen i potencien les pràctiques que van anar bé i es crea un pla de millora de la pràctica d'Scrum a l'equip.

3. **Artefactes:**

(a) *Product backlog:* És una llista de funcionalitats ordenades per prioritat que defineix, amb més o menys detall, com serà el producte final. Aquesta llista és creada i modificada pel PO, tot i que aquest pot demanar ajuda a l'Scrum master per fer-ho.

(b) *Sprint backlog:* És la part del Product backlog que conté les funcionalitats que han sigut assignades a l'esprint actual. Aquestes han d'estar completament definides, i han de contenir la informació necessària per facilitar el seu desenvolupament pel DT.

(c) *Increment:* És el conjunt de funcionalitats del Product backlog que han sigut realitzades. Això implica que han de ser completes, provades, d'alta qualitat i potencialment lliurables.

8.1.3 Control de versions

El control de versions [40] és un sistema que enregistra els canvis realitzats sobre un arxiu o conjunt d'arxius al llarg del temps, de manera que pots recuperar versions específiques en qualsevol moment. Excepte els sistemes de control de versions (Version Control System o VCS en anglès) locals (explicats al següent paràgraf), aquest tipus de producte permet mantenir una còpia del projecte al núvol, accessible des de qualsevol lloc per tots els membres de l'equip autoritzats. Qualsevol tipus d'arxiu pot posar-se sota control de versions. A banda d'això, un sistema de control de versions permet revertir arxius a un estat anterior, revertir un projecte sencer a un estat anterior, comparar canvis al llarg del temps, veure qui va modificar per últim cop alguna cosa que pot causar un error, qui va introduir un error i quan, i algunes utilitats més. Els controladors de versions es poden classificar en 3 tipus diferents depenent d'on es guarda físicament el projecte, es descriuen a continuació:

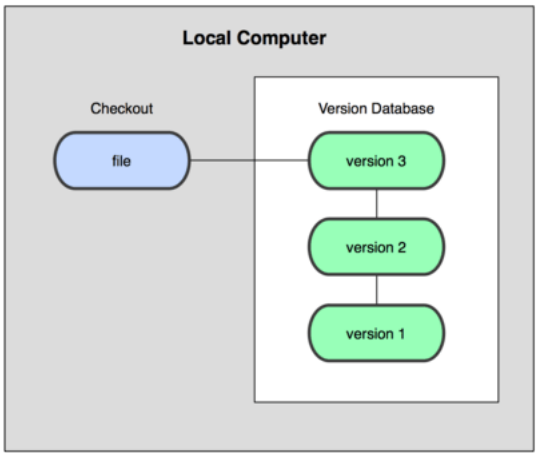
- **VCS local:** Un mètode de control de versions utilitzat per molta gent, tot i que cada vegada menys, és la creació rudimentària de còpies de seguretat de diferents estats del projecte en directoris locals. Aquest enfocament és molt comú donat que és molt simple, però també molt propens a errors. És fàcil oblidar quin és el directori actual i crear, modificar o eliminar algun arxiu d'una versió que no és la desitjada. Per solucionar aquest problema, es va desenvolupar els VCS locals els quals contenien una simple base de dades on enregistraven tots els canvis realitzats dels arxius.

  A la Figura 8.3 podem veure un diagrama que exemplifica l'estructura d'un VCS local. Podem veure com la màquina local conté tant els arxius com l'històric de canvis d'aquest.

- **VCS centralitzat:** Aquest mètode pretén resoldre un problema comú al desenvolupament del software, i és la col·laboració de diferents desenvolupadors d'altres sistemes en un mateix projecte.

  Aquests sistemes tenen un únic servidor que conté tots els arxius versionats, i diversos clients que accedeixen als arxius des d'aquest lloc central. Aquesta configuració ofereix alguns avantatges respecte als

Figura 8.3: Diagrama d'un VCS local.



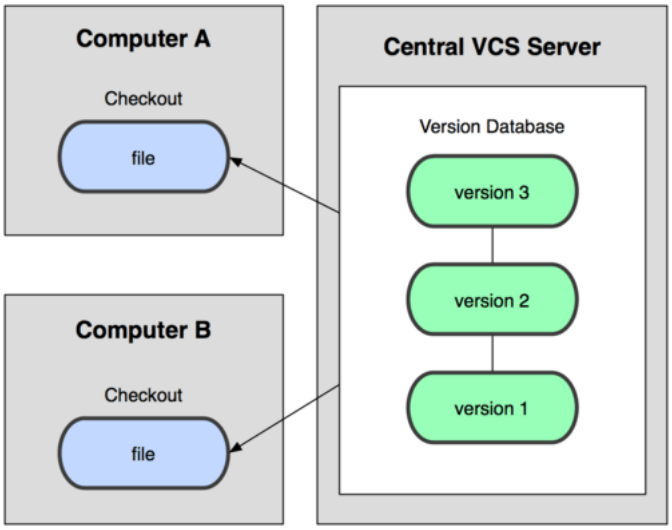
VCS locals. Per exemple, tothom sap (fins a cert punt) en què està treballant la resta de col·laboradors del projecte. A més, és molt més fàcil administrar un únic VCS comú que una base de dades per cada màquina utilitzada.

Per altra banda, també presenta diversos inconvenients. El més important és la concentració de tota la informació en un sol servidor, el que pot portar a perdre tota la feina en cas que aquest falli. O sense ser tan extrems, el servidor pot caure durant un cert període de temps, impedint que l'equip treballi amb normalitat.

A la Figura 8.4 podem veure un diagrama que exemplifica l'estructura d'un VCS centralitzat. Veiem que cada equip conté la seva pròpia còpia de l'arxiu, però que la base de dades de canvis es troba a un servidor central.

- **VCS distribuïts:** Per últim, tenim els sistemes de control de versions distribuïts. Com el seu nom indica, en aquest tipus la informació està distribuïda entre totes les maquines. Això vol dir que, com en els sistemes centralitzats, els arxius es troben en tots els clients, però, la base de dades de canvis es replica tant en el servidor com en la resta d'equips. Això soluciona els problemes de dependència del servidor que presentava els sistemes centralitzats. A més, com que tots els clients

Figura 8.4: Diagrama d'un VCS centralitzat.



contenen la totalitat del repositori, es pot treballar perfectament sense connexió al servidor.

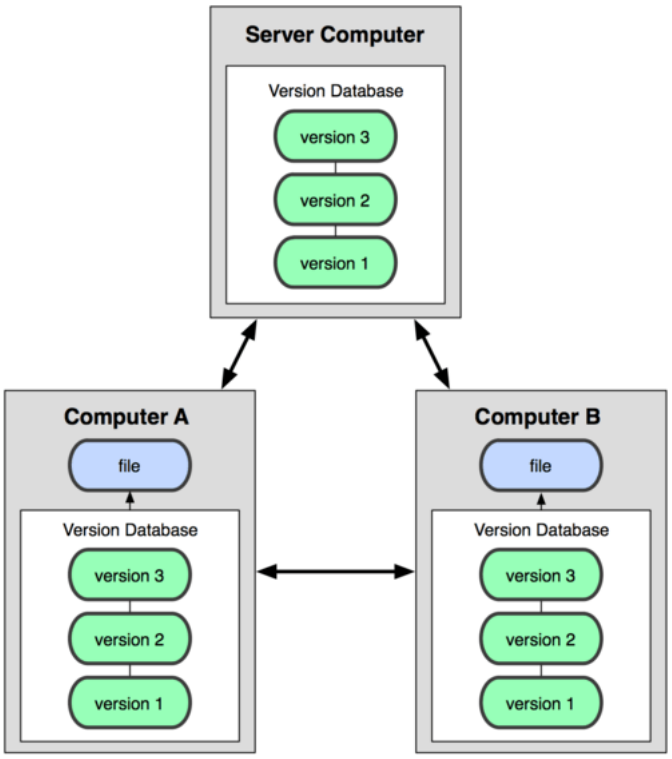
A la Figura 8.5 podem veure un diagrama que exemplifica l'estructura d'un VCS distribuït. Veiem que cada equip conté la seva pròpia còpia de l'arxiu i de la base de dades de canvis, la qual es troba també al servidor central.

Una vegada definit en què consisteix un sistema de control de versions i els seus tipus, parlarem de Git, que és el sistema que nosaltres utilitzem al projecte.

8.1.4 Git

Per començar, repassarem una mica l'origen de Git [41]. El nucli de Linux [42] és un projecte de software de codi lliure amb un abast molt important. Durant la major part del manteniment del nucli de Linux (1991-2002), els canvis en el software van passar a ser pegats i arxius. En 2002,

Figura 8.5: Diagrama d'un VCS distribuït.



el projecte del nucli de Linux va començar a utilitzar un VCS distribuït propietari anomenat BitKeeper [43].

En 2005, la relació entre la comunitat que desenvolupava el nucli de Linux i la companyia que desenvolupava BitKeeper es va trencar, i l'eina va deixar de ser oferida de forma gratuïta. Això va impulsar a la comunitat de desenvolupament de Linux (en particular al seu creador, Linus Torvalds) a desenvolupar la seva pròpia eina basada en l'ús previ de BitKeeper. Els objectius que la nova eina havia de complir van ser els següents:

- Velocitat.
- Disseny simple.
- Suport al desenvolupament paral·lel.
- Distribuït.
- Capacitat de suportar grans projectes (com el nucli de Linux) de manera eficient.

El resultat va ser Git, una eina que cobria tots els seus objectius inicials. A continuació destaquem les característiques més importants de Git [44]:

- **Captures en comptes de canvis:** La principal diferència de Git amb la majoria de VCS és el modelatge que aquest fa de les seves dades. Conceptualment, la majoria de la resta de sistemes emmagatzemen la informació com una llista de canvis als arxius. Altres sistemes, tendeixen a emmagatzemar les dades com els canvis de cada arxiu respecte a la seva versió inicial. Git, en canvi, cada vegada que l'usuari confirma una modificació realitza una captura de l'estat del projecte i guarda una referència a aquesta captura. Per ser eficient, si un arxiu no ha sigut modificat, Git emmagatzema un enllaç a la versió anterior, la qual és idèntica.
- **Localitat de les operacions:** La majoria de les operacions a Git només requereixen arxius i recursos locals per operar. Això elimina la

dependència de connexió a internet, excepte per operacions de compartiment de canvis. A més, també afecta el rendiment del sistema. Com que per moltes operacions no requereix connectar-se a cap servidor extern, pot calcular localment molts dels processos optimitzant recursos.

- **Integritat:** Tota la informació a Git és verificada mitjançant una suma de comprovació abans de ser emmagatzemat, i s'utilitza aquesta suma com a identificador. Això significa que no és possible canviar els continguts de qualsevol arxiu o directori sense que Git sigui capaç de detectar-lo. Aquesta funcionalitat assegura que no es pot perdre informació durant la seva transmissió o patir corrupció d'arxius sense detecció. El mecanisme que usa Git per generar aquesta suma de comprovació es coneix com hash SHA-1 [45]. Es tracta d'una cadena de 40 caràcters hexadecimals (0-9 i A-F), i es calcula basant-se en els continguts de l'arxiu o estructura de directoris.

8.2 Planificació temporal

En aquest apartat s'explica la planificació temporal que segueix el projecte. En primer lloc s'identificaran les tasques que s'han de realitzar per completar-lo i es proposa un pla d'acció per organitzar-les.

8.2.1 Identificació de tasques

Per facilitar la planificació del projecte, aquest ha sigut dividit en 4 parts. Com veurem a continuació, totes les parts no tenen el mateix pes, i per tant, no se li destinen els mateixos recursos. Com que s'empra una metodologia àgil (explicada a l'apartat 8.1), s'ha definit cada tasca com un bloc de feina sense especificar concretament els seus requisits. En els següents paràgrafs, expliquem les parts que conformen el projecte.

La primera part del projecte, anomenada **Fita inicial**, conté les tasques requerides per l'assignatura de GEP. En aquesta assignatura es genera una primera part de la documentació i es fa una breu presentació del

projecte. La feina de GEP s'organitza en 6 lliuraments, que són els que nosaltres hem adoptat com a divisió d'aquest bloc. Les tasques definides són les següents:

1. **Definició de l'abast i la contextualització:** Es tracta de fer una primera reflexió del projecte que es vol realitzar. Aquesta tasca es conforma per 5 apartats diferenciats:
   (a) *Context:* Es realitza una primera contextualització on es defineixen termes i conceptes propis del tema que és objecte d'estudi.
   (b) *Estat de l'art:* Es fa una revisió de la literatura del tema objecte d'estudi: se citen i es resumeixen els resultats d'estudis anteriors deixant clar on és la frontera de coneixement en l'àmbit del projecte. En aquest apartat, també s'estudia si és convenient aprofitar i adaptar una solució existent o si s'ha de dissenyar una de nova.
   (c) *Formulació del problema:* S'especifiquen clarament els objectius del treball i es justifica que el projecte té entitat més que suficient per a ser un TFG.
   (d) *Abast:* Es defineix quin és l'abast del projecte i quins possibles obstacles poden influir-ne.
   (e) *Metodologia i rigor:* Es descriuen els mètodes de treball, les eines de seguiment i el mètode de validació.
2. **Planificació temporal:** En primer lloc, s'especifiquen les tasques, el temps i els recursos. Respecte a aquestes tasques, es descriu la seqüència lògica i s'explicita les seves dependències de precedència. Per altra banda, es proposen solucions a eventuals desviacions, explicant la seva afectació a la duració total del projecte, així com al consum de recursos.
3. **Gestió econòmica i sostenibilitat:** Es descriuen i s'inclouen tots els elements a considerar en l'estimació del pressupost: costos directes per activitat, costos indirectes, amortitzacions, contingències i imprevistos. Una vegada identificats, es fa una estimació de quin pot ser el seu valor final. També es proposa i es descriuen diferents mecanismes de control de desviacions respecte al pressupost. També s'inclou una justificació de la sostenibilitat del projecte en diferents àmbits: econòmic, social i ambiental.

4. **Presentació preliminar:** Es realitza una presentació preliminar de la feina acabada fins al moment per a detectar vicis que calen corregir a l'hora d'efectuar una presentació oral.

5. **Plec de condicions:** Es reflexiona i justifica sobre quines assignatures del grau han aportat més coneixement sobre l'àmbit del projecte. També es justifica la relació del projecte amb les competències tècniques de l'especialitat i el seu nivell d'assoliment.

6. **Document final i presentació oral:** Per acabar aquesta part, es confecciona un document que contingui tot el material redactat després de corregir els errors detectats. Aquest resultat es mostra en una presentació oral on també s'apliquen les correccions dels errors detectats a la presentació preliminar.

El següent bloc de tasques del projecte, anomenat **Plantejament del projecte**, conté el treball previ necessari per dur a terme un desenvolupament adequat. S'analitza detalladament el projecte fixant-se en la idea extreta de la fita inicial i es realitza una immersió en aspectes concrets del treball, tant teòrics com pràctics. Per això s'han definit les següents tasques:

1. **Aprofundiment en l'estat de l'art:** Aquesta ha de ser la primera feina a realitzar. Consisteix a investigar, descobrir i valorar quines tècniques s'utilitzen al mercat actualment. Això permet tenir una visió inicial de quines opcions tenim disponibles, la seva complexitat i els resultats que poden donar.

2. **Selecció d'eines de desenvolupament:** Abans d'iniciar el desenvolupament del projecte, és indispensable valorar les alternatives tecnològiques existents al mercat i decidir quines d'aquestes s'adeqüen més al projecte en qüestió. Aquesta fase pot requerir proves de concepte, si la tecnologia no es coneix anteriorment.

3. **Familiarització amb les dades i l'entorn:** Per últim, és imprescindible conèixer de quines dades disposem i quin és el seu format. Així com familiaritzar-se amb l'entorn seleccionat per tal d'agilitzar i millorar el desenvolupament del producte final.

La tercera part del projecte, anomenada **Desenvolupament del projecte**, conté les tasques corresponents al desenvolupament del producte. Aquesta part és la més important, ja que compren l'assoliment dels objectius principals i, en conseqüència, sol ser a la que més recursos (temporals i econòmics) es destina. Les tasques definides són aquestes:

1. **Preprocés de les dades:** Una vegada conegut quin és el format de les dades, s'ha de reflexionar sobre si aquest és el format idoni o no, i en aquest segon cas, definir quin ho seria. Això pot incloure des de canvis en el tipus de fitxer emprat fins a neteja de dades que no aporten informació però pot alentir el còmput.

2. **Extracció d'informació:** A continuació, hem d'extreure informació de les dades de les quals disposem. Aquesta tasca inclou una estreta col·laboració entre el desenvolupador i l'especialista, ja que és ell qui sap quines dades són més significatives i més útils. L'extracció d'informació consisteix a, a partir d'una gran quantitat de dades difícilment interpretables, donar un conjunt reduït de dades concretes capaces de resumir i destacar la informació útil que aporten.

3. **Representació de la informació:** Per facilitar la interpretació de la informació estreta per part de l'usuari, es generen un conjunt de gràfiques que la mostren d'una forma visual, agradable i funcional. Per això, és necessari seleccionar quin tipus de gràfic s'adapta millor a la informació requerida.

4. **Generació de recomanacions:** Amb l'objectiu d'assistir a l'usuari en el procés de desenvolupament dels inhibidors de la COX-2, implementem la funcionalitat de recomanació de millores. Aquesta consisteix en el fet que la computadora sigui capaç de, analitzant tota la informació extreta prèviament, recomanar a l'especialista quins residus són més favorables o desfavorables per a l'objectiu final: maximitzar la unió entre el lligand i la COX-2.

5. **Desenvolupament d'una aplicació web:** Per últim, volem donar a l'usuari la possibilitat d'interactuar amb el sistema d'una forma senzilla, agradable i funcional. Això es fa amb la finalitat que la utilització d'aquest sistema requereixi el procés mínim d'aprenentatge.

L'últim bloc de tasques del projecte, anomenat **Fita final**, compren la confecció del material que mostrarà els resultats del projecte. Aquest son bàsicament la documentació i la presentació oral. Es divideix en aquestes tasques:

1. **Ultimar la documentació:** S'inclou la informació que manqui i es descriuen les conclusions i els resultats. Després es reestructura tota la documentació de manera que sigui fàcil i practica de llegir. Per últim, es repassa tot el document i s'afegeixen elements extra (figures, taules, gràfics, referències,...) que afavoreixin la correcta comprensió d'aquest.
2. **Material addicional:** Es prepara el material addicional que s'ha d'incloure a la documentació final. Aquest està conformat pel codi i les gràfiques no incloses a cap apartat del document. El material addicional es posiciona al document en forma d'annexos.
3. **Presentació oral:** Una vegada la documentació està completa, es prepara una presentació oral d'entre 20 i 30 minuts on s'explica tot el projecte i es mostren els resultats obtinguts. Això implica crear tot el material de suport necessari.

8.2.2 Pla d'acció

Una vegada s'han definit les tasques que compondran el projecte, podem estimar quina duració en el temps suposa cadascuna i, en conseqüència, quina serà la durada aproximada del projecte.

A la Taula 8.1 es detalla l'estimació temporal de les tasques. Com podem veure, el projecte té una durada aproximada de 485 h de feina. Donat que existeixen algunes limitacions de calendari i de dependència entre tasques, s'ha de realitzar una planificació que asseguri que hi ha temps suficient per enllestir-lo a temps. La data d'inici de projecte és el 15/02/2016, dia en el qual comença el quadrimestre de primavera del grau. La data final és el 28/06/2016, dia en el qual se celebra la presentació final del projecte. A la plana següent (Figura 8.6), podem veure un diagrama de Gantt on es mostra, per cada tasca, les dates de realització i finalització, l'estimació temporal, les dependències amb altres i la precedència.

Taula 8.1: Duració estimada de les tasques del projecte.

Tasques	**Duració (Hores)**
Definició de l'abast i la contextualització	10 h
Planificació temporal	5 h
Gestió econòmica i sostenibilitat	10 h
Presentació preliminar	10 h
Plec de condicions	5 h
Document final i presentació oral	5 h
Aprofundiment en l'estat de l'art	20 h
Selecció d'eines de desenvolupament	30 h
Familiarització amb les dades i l'entorn	30 h
Preprocés de les dades	25 h
Extracció d'informació	30 h
Representació de la informació	80 h
Generació de recomanacions	40 h
Desenvolupament d'una aplicació web	75 h
Ultimar la documentació	30 h
Material addicional	30 h
Presentació oral	50 h
Total	**485 h**

Figura 8.6: Diagrama de Gantt amb la planificació del projecte.

	Nombre	Duración	Inicio	Fin	Predecesoras
1	Definició de l'abast i la contextualització	10h	15/02/2016	17/02/2016	
2	Planificació temporal	5h	17/02/2016	18/02/2016	1
3	Gestió econòmica i sostenibilitat	10h	18/02/2016	21/02/2016	2
4	Presentació preliminar	10h	21/02/2016	23/02/2016	3
5	Plec de condicions	5h	24/02/2016	25/02/2016	1
6	Document final i presentació oral	5h	25/02/2016	26/02/2016	1,2,3,4
7	Aprofundiment en l'estat de l'art	20h	26/02/2016	01/03/2016	1
8	Selecció d'eines de desenvolupament	30h	02/03/2016	09/03/2016	1
9	Familiarització amb les dades i l'entorn	30h	09/03/2016	16/03/2016	8
10	Preprocés de les dades	25h	17/03/2016	23/03/2016	9
11	Extracció d'informació	30h	23/03/2016	30/03/2016	10
12	Representació de la informació	80h	30/03/2016	19/04/2016	11
13	Generació de recomanacions	40h	19/04/2016	28/04/2016	11
14	Desenvolupament d'una aplicació web	75h	29/04/2016	17/05/2016	12,13
15	Ultimar la documentació	30h	17/05/2016	25/05/2016	1,2,3,4,6,7,8,9,10,11,12,13,14
16	Material addicional	30h	25/05/2016	01/06/2016	1,2,3,4,6,7,8,9,10,11,12,13,14
17	Presentació oral	50h	01/06/2016	13/06/2016	15,16

Amb la planificació poden sorgir diferents inconvenients per als quals hem de definir una alternativa que els solucioni. En el nostre cas, tractarem dos imprevistos que poden afectar el transcurs correcte del projecte:

- **Desviacions a la planificació:** El primer d'aquests són les desviacions a la planificació establerta. Això és molt freqüent en un projecte d'una durada considerable com és el cas. Per resoldre aquest aspecte, s'ha intentat sobreestimar temporalment les tasques per poder absorbir el temps extra de les que necessitin més temps amb el temps sobrant d'altres. Com s'utilitzen metodologies àgils amb iteracions, es pot detectar fàcilment endarreriments i solucionar-los a temps.
- **Disponibilitat de l'equip:** Altre problema que podem tenir amb la planificació és la disponibilitat dels participants a l'hora de reunirnos. Això pot allargar els temps d'algunes iteracions, obligant a fer menys o escurçar altres. Per minimitzar aquest inconvenient, és molt recomanable establir les reunions en el mateix moment que es programa cada iteració. Així podem adaptar la feina de cada iteració a la seva duració i tot l'equip adquireix el compromís amb suficient antelació.

8.3 Pressupost i sostenibilitat

En aquest apartat, s'estima el cost aproximat del projecte tenint en compte els recursos utilitzats per cada tasca. Una vegada estimat el cost, es discuteix sobre la sostenibilitat del projecte en diversos aspectes: ambiental, econòmic i social.

8.3.1 Identificació i estimació de costos

Abans d'estimar, cal identificar quins recursos requereix el projecte i quantificar la seva utilització. Els recursos de software i de hardware utilitzats són els següents:

- **Ordinador portàtil Lenovo G510:** Es tracta de l'ordinador portàtil amb el qual es desenvoluparà tot el projecte. És l'únic recurs de hardware que s'utilitza.
- **RStudio [46]:** RStudio és un IDE (Integrated Development Environment) que ofereix una interfície gràfica que facilita la programació amb R [21].
- **PyMOL [26]:** És un sistema de visualització molecular en 3D molt útil per realitzar simulacions d'interaccions moleculars.
- **Shiny [30]:** És un framework per desenvolupar aplicacions web amb R.
- **ShareLaTeX [47]:** És un editor de LaTeX [48] en línia que permet enregistrar projectes LaTeX al núvol.
- **Git (Explicat a l'apartat 8.1.4):** És un sistema de control de versions distribuït.
- **SourceTree [49]:** És una aplicació que ofereix una interfície gràfica per interactuar amb repositoris Git.

A banda d'aquest tipus de recursos, també es requereixen recursos personals. Aquest projecte el realitzarà una única persona en qualitat dels diferents rols requerits:

- **Cap de projecte:** És la persona que té la responsabilitat total del planejament i l'execució fins a l'acabament del projecte.
- **Data scientist:** És la persona especialitzada a extreure informació útil de grans conjunts de dades.
- **Desenvolupador de software:** És la persona encarregada de dissenyar i implementar el software que conforma el projecte.

Per últim, s'han de tenir en compte els recursos indirectes. Els recursos indirectes són els que deriven de la utilització d'uns altres. Aquests recursos són els següents:

- **Electricitat:** L'electricitat és necessària com a font de l'energia per al PC.
- **Paper:** El paper és necessari a l'hora d'imprimir la documentació.
- **Tinta:** La tinta és necessària l'hora d'imprimir la documentació.

A les Taules 8.2, 8.3, 8.4, 8.5, 8.6, 8.7 i 8.8 s'identifiquen quins recursos són necessaris per realitzar les tasques definides a l'apartat de planificació. S'han dividit segons el bloc de tasques al qual pertanyen per facilitar la lectura de les taules. Les tasques són les utilitzades en el diagrama de Gantt. Per simplificar les taules, no incloem l'electricitat, ja que va lligada a l'ús del PC.

Quan ja sabem quins recursos i quina quantitat d'ells s'utilitza en cada tasca, podem estimar el cost del projecte. Aquestes estimacions les dividirem en 4 taules segons el tipus de recurs del qual es parli. La Taula 8.9 correspon als recursos de personal, la 8.10 al hardware, la 8.11 al software i la 8.12 als indirectes.

Per últim, ajuntem tots els costos i calculem el cost total del desenvolupament del projecte. Podem veure el resultat a la Taula 8.13

El cost total estimat és de 14.788 €. Com en tota estimació, existeix una possible variació en la duració total de les tasques que componen el projecte. Aquestes variacions afecten directament al cost total del projecte. Per evitar aquest problema, s'ha intentat sobreestimar la duració de les tasques per assegurar que el projecte es pot acabar en aquest temps. En el cas que sobri temps, sempre es pot intentar millorar la feina o implementar alguna petita funcionalitat extra.

A següent apartat, analitzarem si el projecte és sostenible econòmicament, així com socialment i ambientalment.

8.3.2 Sostenibilitat

Per analitzar la sostenibilitat del projecte, s'ha seguit el model proposat a l'assignatura de GEP. Aquest consisteix a donar resposta a un conjunt

Taula 8.2: Relació entre les tasques de la fita inicial del projecte i els recursos que utilitzen (Part 1).

Tasca	**Recurs**	**Hores**
Definició de l'abast i la contextualització		
	Cap de projecte	10 h
	ShareLaTeX	10 h
	PC Lenovo G510	10 h
Planificació temporal		
	Cap de projecte	5 h
	ShareLaTeX	5 h
	PC Lenovo G510	5 h
Gestió econòmica i sostenibilitat		
	Cap de projecte	10 h
	ShareLaTeX	10 h
	PC Lenovo G510	10 h
Presentació preliminar		
	Cap de projecte	10 h
	ShareLaTeX	10 h
	PC Lenovo G510	10 h

Taula 8.3: Relació entre les tasques de la fita inicial del projecte i els recursos que utilitzen (Part 2).

Tasca	Recurs	Hores
Plec de condicions		
	Cap de projecte	5 h
	ShareLaTeX	5 h
	PC Lenovo G510	5 h
Document final i presentació oral		
	Cap de projecte	5 h
	ShareLaTeX	5 h
	PC Lenovo G510	5 h

de preguntes que forcen a reflexionar sobre aquest aspecte. En primer lloc, parlarem de la sostenibilitat econòmica del projecte:

- *Existeix una avaluació de costos, tant de recursos materials com humans?* Si, a l'apartat 8.3.1 del present document podem trobar l'anàlisi dels costos del projecte. En ell es contemplen els costos humans, de software, de hardware i indirectes.
- *S'ha tingut en compte el cost dels ajustaments / actualitzacions / reparacions durant la vida útil del projecte?* En aquest projecte no es contempla una fase d'ajustament ni un procés d'actualització futura. En el cas de les reparacions, es considera que els costos són negligibles donada l'absència de costos de software, les petites dimensions del pressupost de hardware i la simplicitat del projecte que repercuteix en menys hores de feina en aquest procés.
- *El cost del projecte ho faria viable si hagués de ser competitiu?* Si, donat que els costos de software i hardware són mínims, el cost total el conformen bàsicament els de personal i els indirectes. Això fa que el

Taula 8.4: Relació entre les tasques del plantejament del projecte i els recursos que utilitzen.

Tasca	**Recurs**	**Hores**
Aprofundiment en l'estat de l'art		
	Desenvolupador de software	20 h
	PC Lenovo G510	20 h
Selecció d'eines de desenvolupament		
	Desenvolupador de software	30 h
	RStudio	30 h
	PC Lenovo G510	30 h
Familiarització amb les dades i l'entorn		
	Desenvolupador de software	30 h
	RStudio	30 h
	PC Lenovo G510	30 h

Taula 8.5: Relació entre les tasques del desenvolupament del projecte i els recursos que utilitzen (Part 1).

Tasca	**Recurs**	**Hores**
Preprocés de les dades		
	Desenvolupador de software	15 h
	Data scientist	10 h
	RStudio	25 h
	Git	25 h
	SourceTree	25 h
	PC Lenovo G510	25 h
Extracció d'informació		
	Desenvolupador de software	10 h
	Data scientist	20 h
	RStudio	30 h
	Git	30 h
	SourceTree	30 h
	PC Lenovo G510	30 h

Taula 8.6: Relació entre les tasques del desenvolupament del projecte i els recursos que utilitzen (Part 2).

Tasca	**Recurs**	**Hores**
Representació de la informació		
	Desenvolupador de software	50 h
	Data scientist	30 h
	RStudio	80 h
	Git	80 h
	SourceTree	80 h
	PC Lenovo G510	80 h
Generació de recomanacions		
	Desenvolupador de software	30 h
	Data scientist	10 h
	RStudio	40 h
	Git	40 h
	SourceTree	40 h
	PyMOL	10 h
	PC Lenovo G510	40 h

Taula 8.7: Relació entre les tasques del desenvolupament del projecte i els recursos que utilitzen (Part 3).

Tasca	**Recurs**	**Hores**
Desenvolupament d'una aplicació web		
	Desenvolupador de software	75 h
	RStudio	75 h
	Git	75 h
	SourceTree	75 h
	Shiny	75 h
	PC Lenovo G510	75 h

preu s'ajusti el màxim al desenvolupament del producte i, per tant, el fa competitiu envers altres productes similars.

- *Es podria realitzar un projecte similar en molt menys temps o amb molts menys recursos i, per tant, menor cost?* Probablement si es podria desenvolupar en un temps menor, s'ha de tenir en compte que el projecte ha sigut desenvolupat per un estudiant de grau en enginyeria informàtica a l'últim any del grau. Això vol dir que un professional, amb experiència prèvia en projectes del mateix àmbit realitzaria un procés més àgil i directe. De totes maneres, a l'àmbit econòmic seria molt difícil reduir costos, ja que, encara que sigui possible disminuir el temps, seria necessària la contractació de personal més qualificat, el que sovint implica uns honoraris més elevats.

- *El temps dedicat a cada tasca és proporcional a la seva importància (s'ha dedicat molt de temps a desenvolupar parts del projecte que podien haver estat reutilitza-des de tecnologies / projectes / coneixements existents)?* La duració de cada tasca s'ha adequat al valor que aportava al producte final. Quant a la reutilització de feina existent, s'ha tractat de minimitzar la implementació de software que es pogués resoldre amb alguna eina existent.

Taula 8.8: Relació entre les tasques de la fita final del projecte i els recursos que utilitzen.

Tasca	**Recurs**	**Hores**
Ultimar la documentació		
	Cap de projecte	30 h
	ShareLaTeX	30 h
	Paper	-
	Tinta	-
	PC Lenovo G510	30 h
Material addicional		
	Desenvolupador de software	30 h
	PC Lenovo G510	30 h
Presentació oral		
	Cap de projecte	50 h
	ShareLaTeX	50 h
	PC Lenovo G510	50 h

Taula 8.9: Estimació del cost dels recursos de personal.

Rol	**Hores**	**Cost per hora**	**Cost total**
Cap de projecte	125 h	40 €/h	5.000 €
Data scientist	70 h	35 €/h	2.450 €
Desenvolupador	290 h	25 €/h	7.250 €
Total	**485 h**		**14.700 €**

Taula 8.10: Estimació del cost dels recursos de hardware.

Producte	**Preu**	**Hores**	**Vida útil**	**Amortització**
PC Lenovo G510	500 €	485 h	2 anys	58 €
Total	**500 €**			**58 €**

Taula 8.11: Estimació del cost dels recursos de software.

Producte	**Preu**	**Hores**	**Vida útil**	**Amortització**
RStudio	0 €	310 h	2 anys	0 €
ShareLaTeX	0 €	125 h	2 anys	0 €
Git	0 €	250 h	2 anys	0 €
SourceTree	0 €	250 h	2 anys	0 €
Shiny	0 €	75 h	2 anys	0 €
Pymol	0 €	10 h	2 anys	0 €
Total	**0 €**			**0 €**

Taula 8.12: Estimació del cost dels recursos indirectes.

Producte	**Preu**	**Unitats**	**Cost aproximat**
Electricitat	0,1275 €/kWh	48,5 kWh	6 €
Paper	4 €	1	4 €
Tinta	20 €	1	20 €
Total			**30 €**

Taula 8.13: Estimació del cost total del projecte.

Concepte	**Cost aproximat**
Costos de personal	14.700 €
Costos de hardware	58 €
Costos de software	0 €
Costos indirectes	30 €
Total	**14.788 €**

- *Està prevista o hi ha col·laboració amb algun altre projecte (acadèmic, empresa, associació, etc.)?* El producte fa servir les dades extretes mitjançant un projecte previ. De la mateixa forma, s'espera que en un futur es realitzi un altre projecte capaç de fer evolucionar l'eina resultant d'aquest.

En general, el projecte demostra una sostenibilitat força favorable quan ens referim a l'àmbit econòmic. Pot ser que la consideració dels costos de reparacions com a negligibles afecti negativament a l'estimació total, però mai en gran mesura, per les raons que s'han exposat. Per altra banda, veiem que seria possible reduir el temps de desenvolupament, tot i que això no implicaria una reducció dels costos econòmics. A continuació, fem l'anàlisi dels aspectes de caràcter social:

- *Quina és la situació social i política del país / lloc / ciutat / ... on realitzaràs el teu projecte? I la del sector a què inclou el teu projecte?* Espanya és un país amb un dels millors sistemes sanitaris públics del món. En l'actualitat, el país es troba en una situació de crisi econòmica que perdura des de l'any 2008. Aquesta situació ha repercutit en el seu sistema sanitari, ja que s'han reduït els pressupostos destinats a aquest servei. Més concretament, l'àmbit de la investigació en general, i farmacèutic en particular, ha vist reduït el suport governamental que rebien anteriorment. Per altra banda, s'ha instaurat un sistema de co-pagament de fàrmacs amb el qual els pacients han d'abonar una part del

tractament que segueixen (variant depenent de la situació econòmica de l'usuari i de la tipologia del tractament).

- *Creus que la teva activitat podria afavorir o empitjorar aquesta situació?* El projecte que s'ha desenvolupat podria ajudar en la millora de qualitat dels fàrmacs antiinflamatoris inhibidors de la COX-2. A més, també abaratiria el cost del procés de disseny del fàrmac. Això desencadena en una major efectivitat dels tractaments, i en una reducció en el cost d'aquest. La primera conseqüència, afecta directament en un estalvi en el servei i en prestacions per baixes laborals. La segona, pot suposar un estalvi al pressupost destinat a la compra de fàrmacs per part de la seguretat social.

- *Hi ha una necessitat real del teu producte / servei?* La millora dels fàrmacs existents i la millora del procés de disseny dels nous són fonamentals per la millora de l'esperança de vida de l'espècie. Tot i que aquest projecte només afectaria un tipus concret de fàrmacs, aportaria a l'avanç en aquesta direcció.

- *Satisfer aquesta necessitat millora la qualitat de vida dels consumidors?* Com ja hem comentat a les preguntes anteriors, la qualitat de vida dels usuaris es veuria positivament afectada per la millora dels tractaments farmacèutics. Indirectament, un estalvi en els costos del sistema sanitari afavoreix la inversió en altres camps, el que permet més millores.

- *El resultat del projecte, En què / Com canviarà la vida de l'usuari?* Reduint els temps de recuperació d'algunes malalties i obtenint un servei sanitari de major qualitat gràcies als estalvis a la compra de fàrmacs.

- *Hi ha algun col·lectiu que es vegi perjudicat pel TFG, i en quina mesura?* A l'hipotètic cas en el qual es desenvolupés un nou fàrmac antiinflamatori que superes l'efectivitat dels utilitzats actualment o que almenys la mantingues, però amb una reducció de costos, les empreses que comercialitzen les opcions que monopolitzen el mercat es veurien perjudicades. De totes maneres, això forçaria a aquestes corporacions a destinar més recursos per dissenyar nous productes més efectius i/o econòmics.

Com podem veure, el projecte és molt beneficiós per a la societat, no només del territori on se situa, sinó per tot el món. L'únic col·lectiu afectat

serien les companyies farmacèutiques que comercialitzen els fàrmacs utilitzats actualment, però això comportaria una situació en la qual la societat tornaria a veure's afavorida. Per últim, parlarem de la sostenibilitat mediambiental del projecte. Donat que el projecte consisteix en el desenvolupament d'un producte software, s'ha decidit eliminar les preguntes relacionades amb la manufacturació d'aquest. A continuació es responen les preguntes adients:

- *Quins recursos es necessitaran en les diferents fases del projecte?* Donat que estem comentant l'àmbit ambiental, no inclourem els recursos personals ni de software, ja que no generen cap tipus d'impacte ambiental. Una vegada dit això, l'únic recurs utilitzat durant el desenvolupament del projecte és un ordinador portàtil: PC Lenovo G510.

- *Què consum tindran aquests recursos durant el desenvolupament del projecte i posteriorment durant la seva posada en marxa i vida útil? Quin és l'impacte ambiental d'aquest consum (mesurat en tones de CO2, per exemple?)* El PC només consumeix recursos durant la seva utilització, ja que la seva fabricació no s'atribueix al desenvolupament del projecte. Per a totes les fases, el consum elèctric del PC és de 100 W. El consum total s'ha de calcular tenint en compte les hores totals d'ús. Sabem que la durada aproximada del desenvolupament és de 485 hores, el que comporta un consum de 48,5 kWh. El projecte no té fase de posada en marxa i per tant no consumeix cap recurs. Dit això, no es pot estimar el consum durant la seva vida útil, ja que dependrà de les hores en les quals la màquina que l'executi es mantingui encesa. Per facilitar una visió de l'impacte ambiental del consum elèctric del PC, calcularem les emissions per hora d'ús de l'ordinador. Segons la OCCC (Oficina Catalana del Canvi Climàtic) el mix elèctric de 2015 (emissions de CO2 associades a la producció d'un kWh) va ser de 302 g CO2/kWh [50]. Donat que el consum de l'ordinador és de 0,1kW (100W), les seves emissions per hora són de 30,2 g CO2. El que implica que l'impacte ambiental del desenvolupament del projecte és de 14,6 kg CO2.

- *Quin consum i impacte ambiental tindria realitzar la mateixa activitat sense l'existència del teu TFG (estalvi de paper i altres materials i/o energia?)* El consum seria el mateix, ja que no s'ha tingut en compte l'impacte ambiental del paper ni la tinta utilitzats.

Taula 8.14: Matriu de sostenibilitat del projecte.

	PPP	**Vida útil**	**Riscos**
Ambiental	10	20	0
Econòmic	9	20	-3
Social	10	20	-5

- *Quins recursos poden reaprofitar d'altres projectes?* Cap, el projecte és autocontingut i utilitza els seus propis recursos.
- *Durant el desenvolupament del teu producte es generarà algun tipus de contaminació?* Si, la derivada de la producció de l'energia elèctrica que consumeix l'ordinador.
- *Amb la implantació del projecte s'augmenta o es disminueix la petjada ecològica?* Donat que el producte pretén reduir la durada del procés de disseny dels fàrmacs antiinflamatoris inhibidors de la COX-2, es redueix la petjada ecològica que es generaria durant aquest temps de treball estalviat.
- *Quines parts del projecte podran reciclar-se o reutilitzar-se en altres projectes?* Tot el projecte. A partir d'aquest pot originar-se'n altres que el continuïn i el millorin.

Per finalitzar, s'ha confeccionat la matriu de sostenibilitat que pretén resumir tota la informació analitzada durant aquest apartat. La matriu s'ha dissenyat seguint la documentació de l'assignatura de GEP i s'ha emplenat d'acord amb les respostes anteriors. Podem veure la matriu a la Taula 8.14.

9 Bibliografia

[1] Youssef Harrak†, Giovanni Casula§, Joan Basset†, Glòria Rosell†, Salvatore Plescia§, Demetrio Raffa§, Maria Grazia Cusimano§, Ramon Pouplana*‡ i Maria Dolors Pujol*†, *Synthesis, Anti-Inflammatory Activity, and in Vitro Antitumor Effect of a Novel Class of Cyclooxygenase Inhibitors: 4-(Aryloyl)phenyl Methyl Sulfones*, † Laboratori de Química Farmacèutica (Unitat Associada al CSIC), ‡ Laboratori de Físico-Química Facultat de Farmàcia, Universitat de Barcelona, Av. Diagonal 643, E-08028 Barcelona, Spain. § Dipartamento di Chimica e Tecnologie Farmaceutiche, Facoltà di Farmacia, Università degli Studi di Palermo, Via Archirafi, 32-90123 Palermo, Italy. J. Med. Chem., 2010, 53 (18), pp 6560–6571 DOI: 10.1021/jm100398z

[2] *Facultat d'Informàtica de Barcelona*, http://www.fib.upc.edu/.

[3] *Universitat Politècnica de Catalunya*, http://www.upc.edu/.

[4] Salido, M.†, Abásolo,L.† i Bañares, A.‡, *Revisión de los antiinflamatorios inhibidores selectivos de la ciclooxigenasa-2*, † Médico Interno Residente, ‡ Facultativo Especialista de Área. Sº de Reumatología. Hospital Clínico San Carlos. Madrid. Información Terapéutica del Sistema Nacional de Salud, Vol. 25–N.º2-2001

[5] R.M. Botting, *Inhibitors of Cyclooxygenases: mechanisms, selectivity and uses*, The William Harvey Research Institute, The John Vane Science Centre, St Bartholomew's and the London School of Medicine and Dentistry, Queen Mary, University of London, U.K. Journal of physiology and pharmacology 2006, 57, Supp 5, 113124

[6] *Fundación Pública Andaluza para la Investigación Biosanitaria en Andalucía Oriental*, http://www.fibao.es/.

[7] *Portal de Medicina Molecular de FIBAO - Glosario*, http://medmol.es/glosario/.

[8] Jürgen Drews, *Drug Discovery: A Historical Perspective*, Science 17 Mar 2000: Vol. 287, Issue 5460, pp. 1960-1964 DOI: 10.1126/science.287.5460.1960

[9] *Jones Research Group*, https://joneslab.chem.wsu.edu/.

[10] *Washington State University*, https://wsu.edu/.

[11] *Diseño molecular de nuevos fármacos ayudado por computadora*, Introducción al modelado molecular, Instituto de Ciencias Físicas (ICF) de la Universidad Nacional Autónoma de Mexico (UNAM), 25 de febrero de 2011.

[12] Liang-Cheng Tu i Jun Luo, *Experimental tests of Coulomb's Law and the photon rest mass*, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074. 5 January 2004.

[13] Jan Hauke i Tomasz Kossowski, *Comparison of values of Pearson's and Spearman's correlation coeficients on the same sets of data*, Adam Mickiewicz University, Institute of Socio-Economic Geography and Spatial Management, Poznań, Poland, April 19, 2011.

[14] William Mendenhall†, Robert J. Beaver‡ i Barbara M. Beaver‡, *Introducción a la probabilidad y estadística*, †University of Florida, ‡University of California. 2010, ISBN 9786074814668

[15] J.C. Escalona, R. Carrasco i J. A. Padrón, *Introducción al diseño de Fármacos*. Folleto para la docencia de la asignatura de Farmacia, Universidad de Oriente, 03 de marzo de 2005.

[16] *Protein Data Bank*, http://www.rcsb.org/pdb/home/home.do.

[17] Stefan H. Unger, *Consequences of the Hansch Paradigm for the Pharmaceutical Industry*, Medicinal Chemistry. A series of monographs. Volume 11-IX, chapter 2. Department of pharmacology, University of Nijmegen. Nijmegen, the Netherlands. 2014, ISBN 9781483294834.

[18] Cramer, R. D.; Patterson, D. E. i Bunce, J. D., *Comparative molecular field analysis (CoMFA)*, 1. Effect of shape on binding of steroids to carrier proteins." Journal of the American Chemical Society 110.18 (1988): 5959-5967.

[19] *RapidMiner*, https://rapidminer.com/.

[20] *Weka*, http://www.cs.waikato.ac.nz/ml/weka/.

[21] *R*, https://www.r-project.org/.

[22] *Python*, https://www.python.org/.

[23] *Matlab*, http://es.mathworks.com/products/matlab/.

[24] *Hohli*, http://hohli.com/.

[25] *ChartGo*, http://www.chartgo.com/.

[26] *Pymol*, https://www.pymol.org/.

[27] *Jmol*, http://jmol.sourceforge.net/.

[28] *Django*, https://www.djangoproject.com/.

[29] *Spring*, https://spring.io/.

[30] *Shiny*, http://shiny.rstudio.com/.

[31] *JavaScript*, https://www.javascript.com/.

[32] *Bio3D*, http://thegrantlab.org/bio3d/index.php.

[33] David Whitford, *Proteins: Structure and Function.* 2005, ISBN 9780471498940.

[34] Carlos Blé Jurado, *Diseño Ágil con TDD*, 2010, ISBN 978-1445264714

[35] Kent Beck, *Extreme Programming Explained: Embrace Change.*, Addison-Wesley, 1999, ISBN 978-0321278654

[36] Kent Beck, *Test-Driven Development: By Example*, Addison-Wesley, 2002, ISBN 978-0321146533

[37] *Agile-Spain*, http://agile-spain.org/utiles/manifesto-agil/.

[38] Ken Schwaber, *Agile Project Management with Scrum*, Microsoft Press, 2009, ISBN 978-0735637900

[39] Hirotaka Takeuchi i Ikujiro Nonaka, *The New New Product Development Game*, Harvard Business Review, January 1986.

[40] *Git, Documentación 1.1: Empezando - Acerca del control de versiones*, https://git-scm.com/book/es/v1/Empezando-Acerca-del-control-de-versiones.

[41] *Git, Documentación 1.2: Empezando - Una breve historia de Git*, https://git-scm.com/book/es/v1/Empezando-Una-breve-historia-de-Git.

[42] *Linux - What is Linux?*, https://www.linux.com/what-is-linux.

[43] *BitKeeper - About Us*, http://www.bitkeeper.com/company_about_us.

[44] *Git, Documentación 1.3: Empezando - Fundamentos de Git*, https://git-scm.com/book/es/v1/Empezando-Fundamentos-de-Git.

[45] Christophe De Cannière i Christian Rechberger, *Finding SHA-1 Characteristics: General Results and Applications*. Advances in Cryptology – ASIACRYPT 2006.

[46] *RStudio*, https://www.rstudio.com/.

[47] *ShareLaTeX - About*, https://es.sharelatex.com/about.

[48] *LaTeX*, https://www.latex-project.org/.

[49] *SourceTree*, https://www.sourcetreeapp.com/.

[50] OCCC (Oficina Catalana del Cambio Climatico), *Nota informativa sobre la metodologia de estimación del mix elèctrico por parte de la oficina catalana del cambio climático (OCCC)*. 19 de febrero de 2016.

10 Annex

Figura 10.1: Gràfica de perfils energètics globals per lligand referent a la primera descomposició de l'energia total.

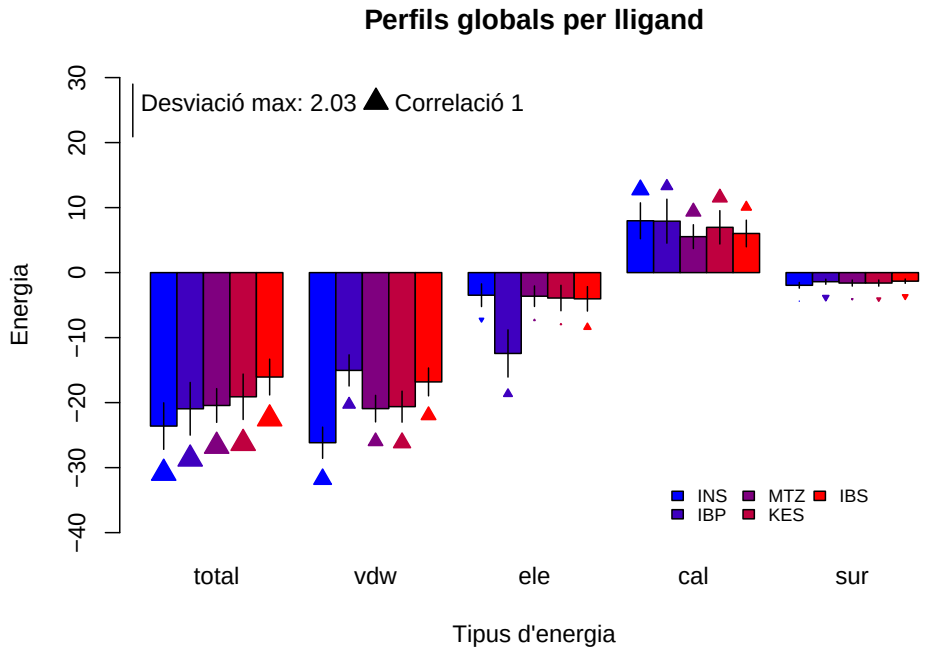


Figura 10.2: Gràfica de perfils energètics globals per lligand referent a la segona descomposició de l'energia total.

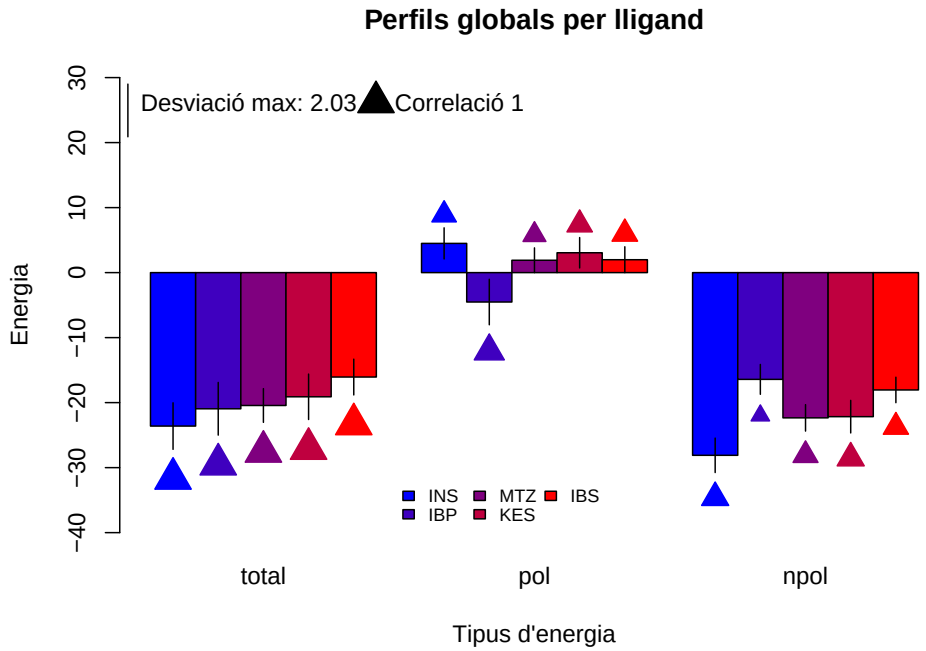


Figura 10.3: Gràfica de perfils energètics globals per lligand referent a la tercera descomposició de l'energia total.

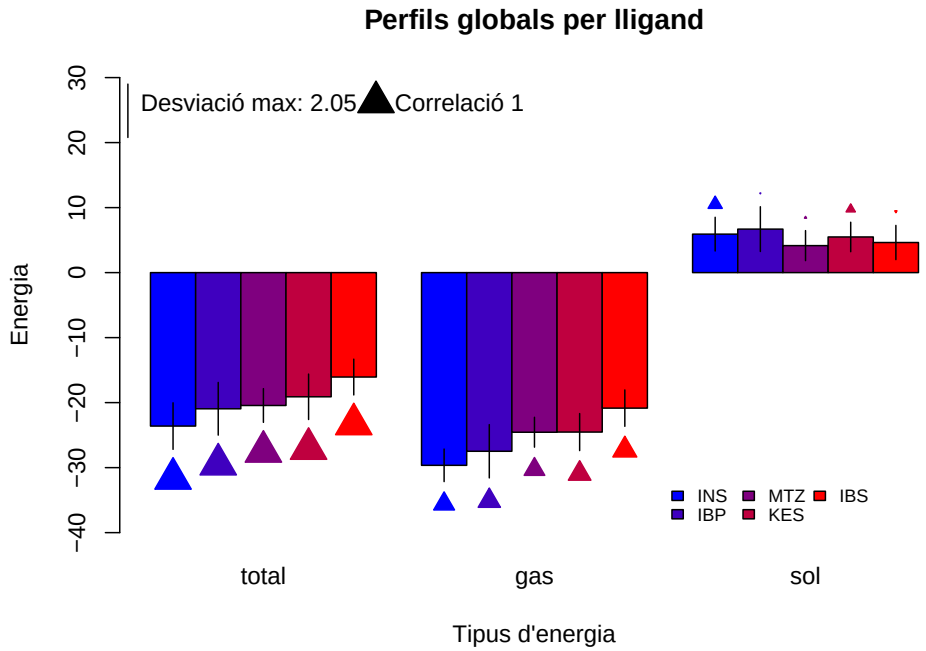


Figura 10.4: Gràfica de perfils energètics per residu referent a l'energia electrostàtica.

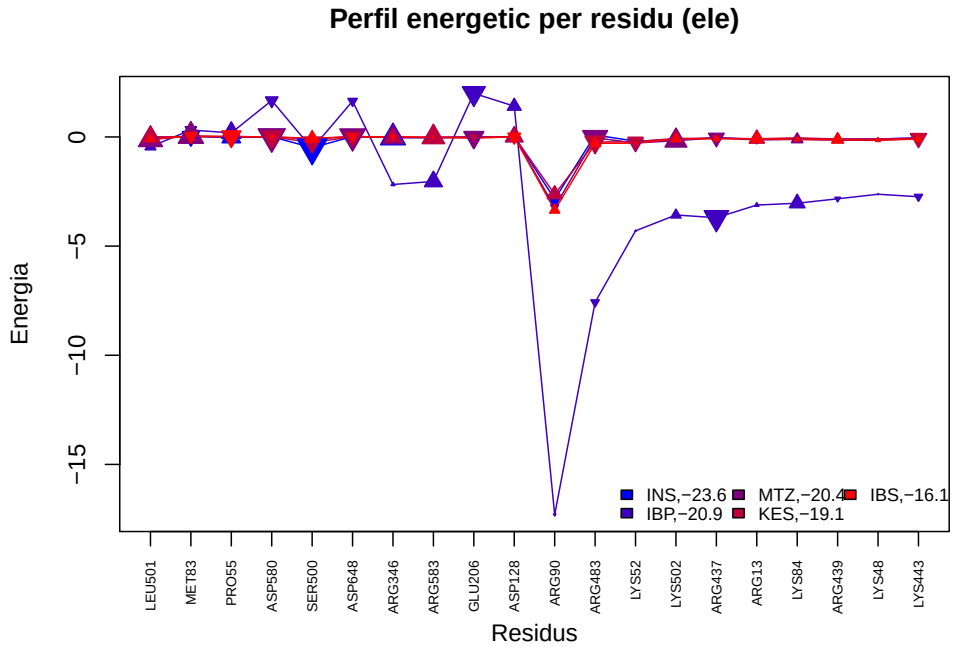


Figura 10.5: Gràfica de perfils energètics per residu referent a l'energia de Van der Waals.

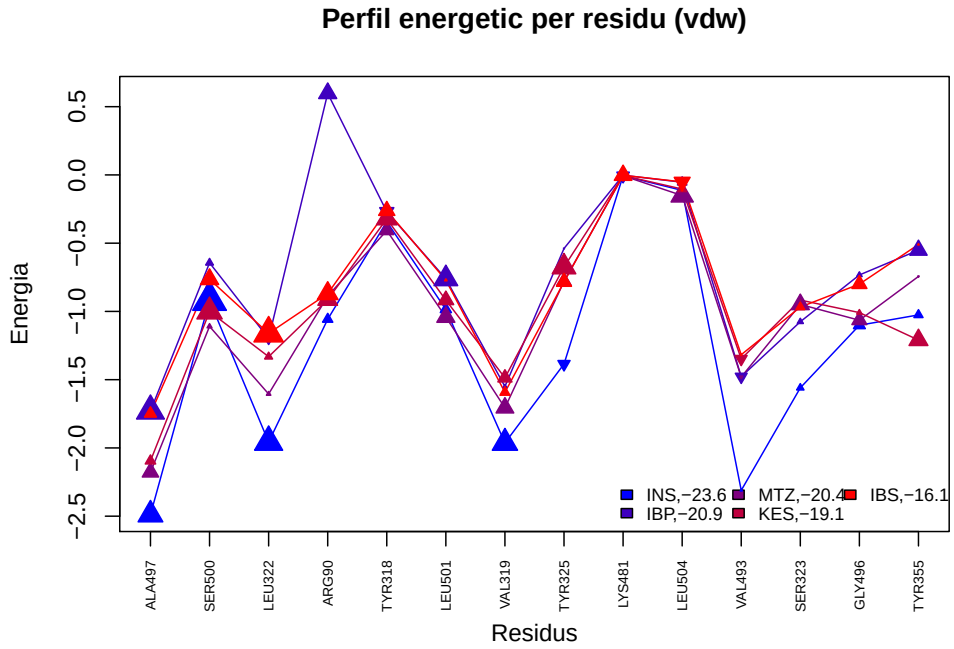


Figura 10.6: Gràfica de perfils energètics per residu referent a l'energia de solvatació polar.

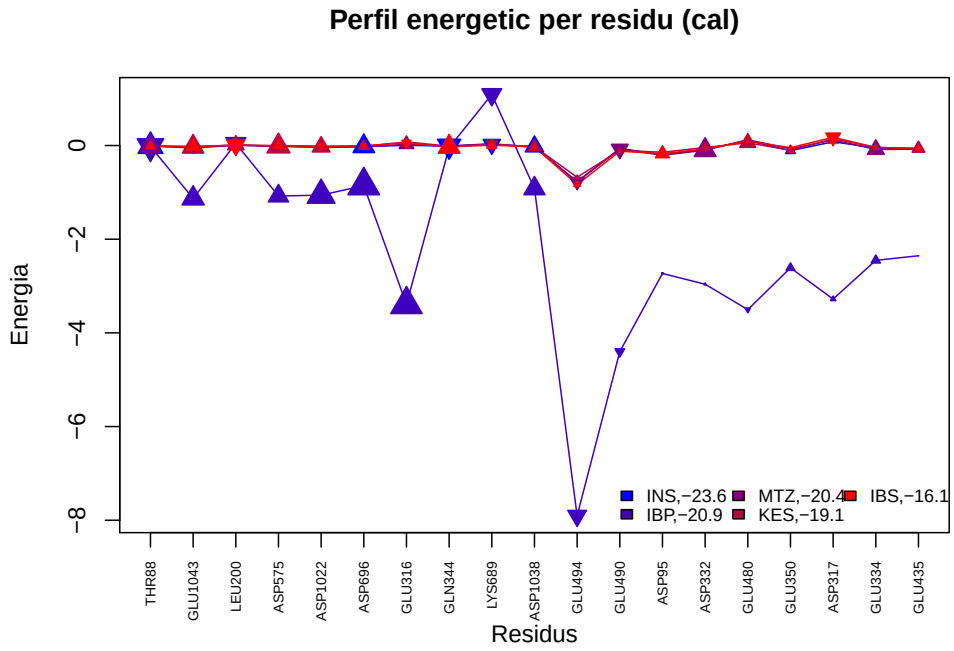


Figura 10.7: Gràfica de perfils energètics per residu referent a l'energia de solvatació apolar.

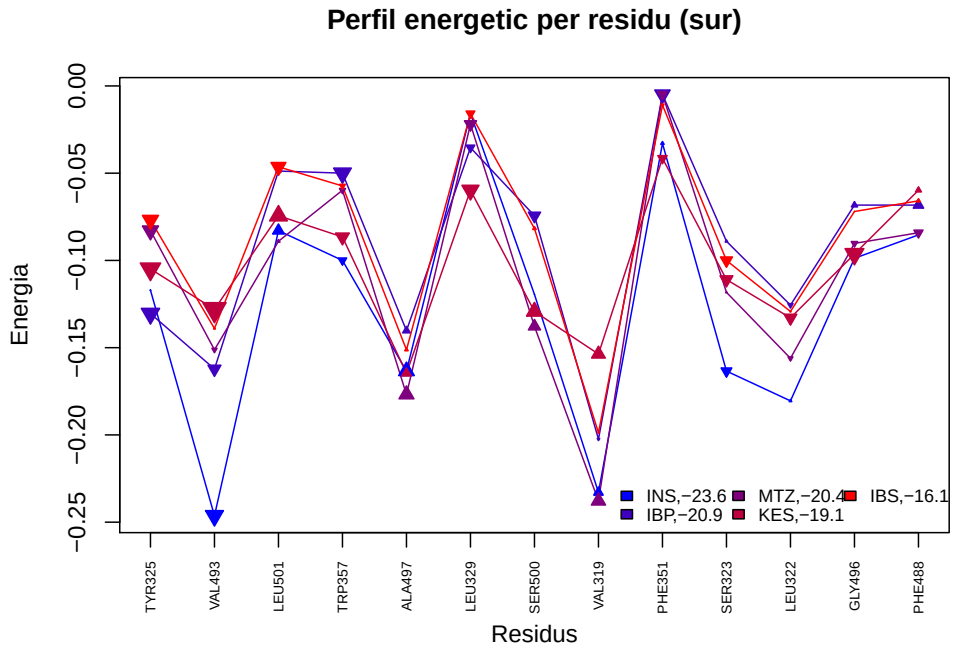


Figura 10.8: Gràfica de perfils energètics per residu referent a l'energia en gas.

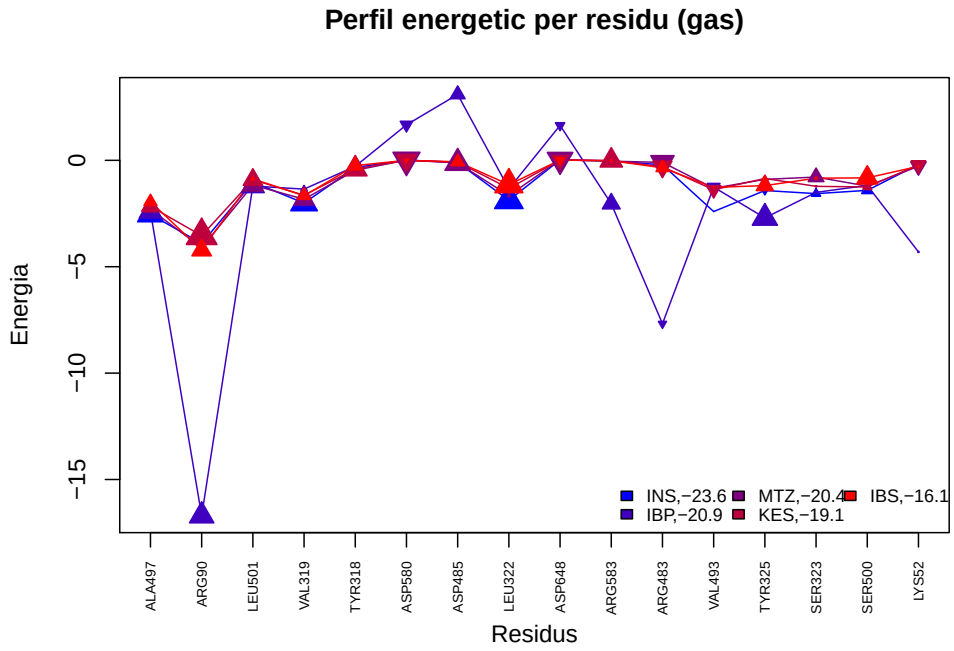


Figura 10.9: Gràfica de perfils energètics per residu referent a l'energia de solvatació.

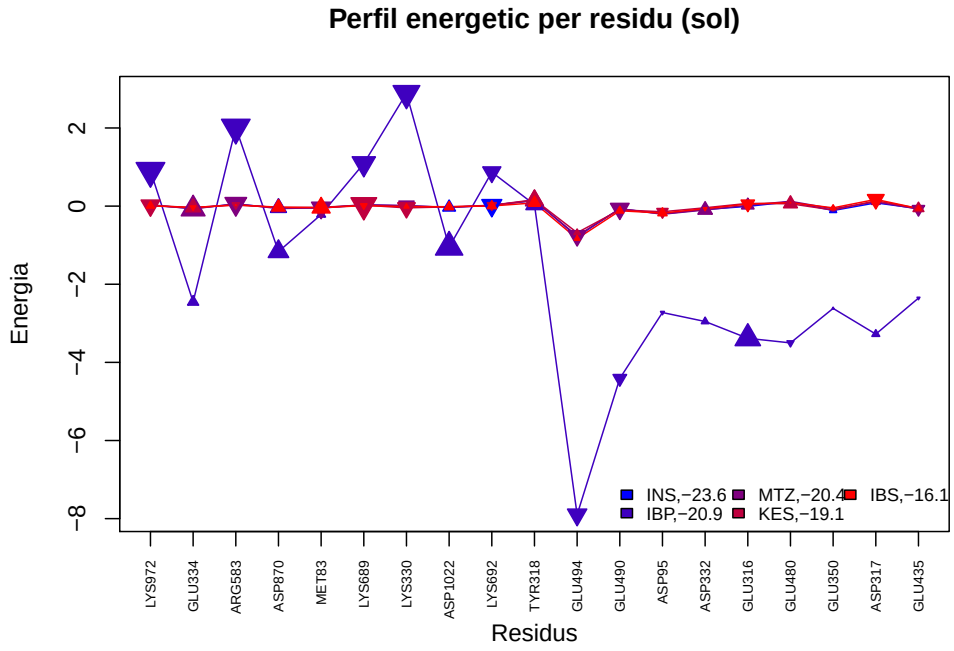


Figura 10.10: Gràfica de perfils energètics per residu referent a l'energia polar.

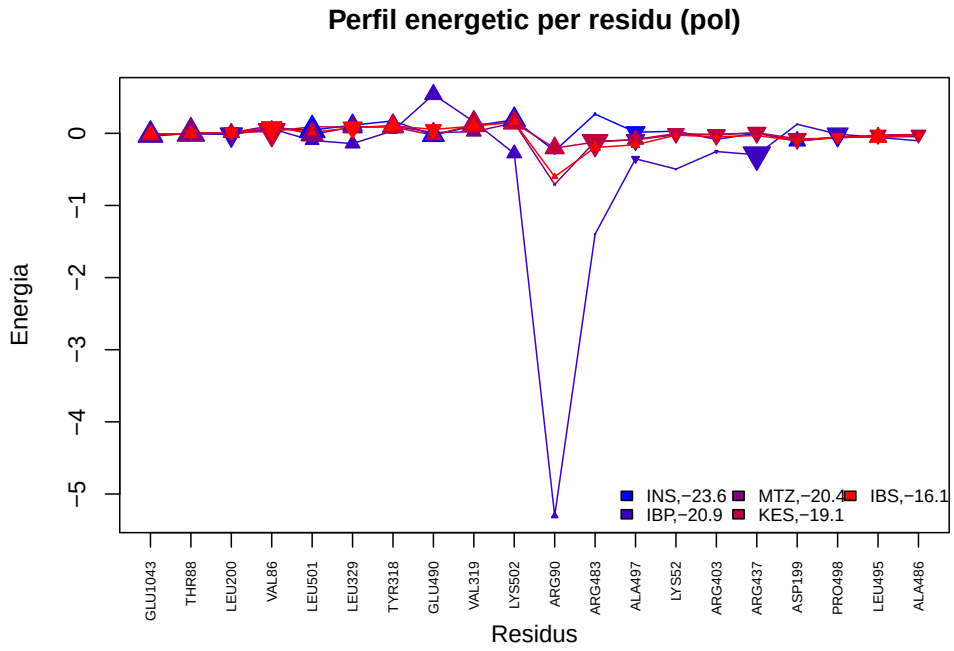


Figura 10.11: Gràfica de perfils energètics per residu referent a l'energia apolar.

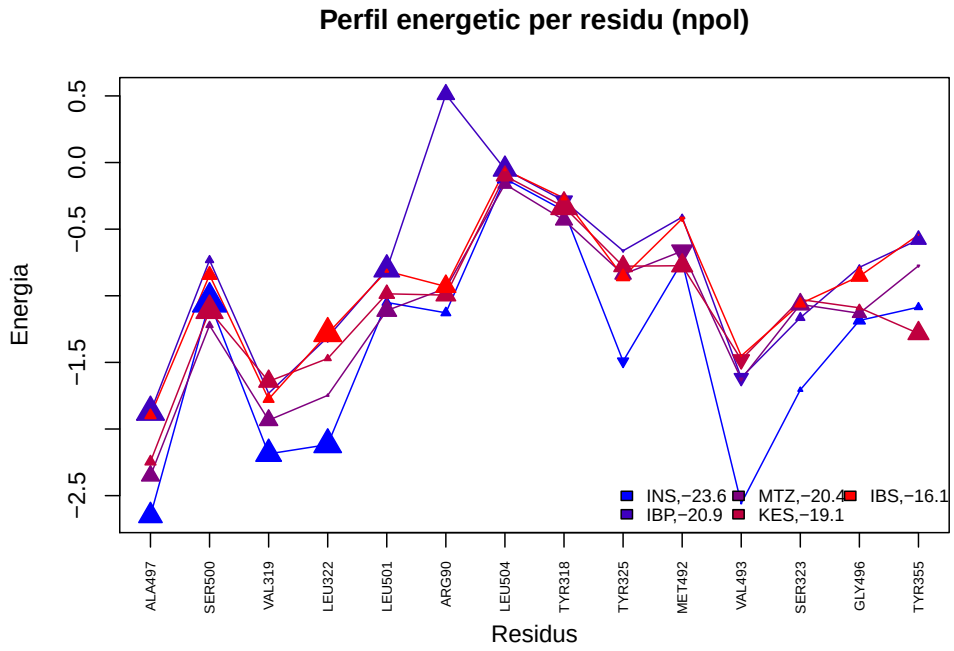


Figura 10.12: Aspecte de l'aplicació web a la vista de perfils globals per lligand.

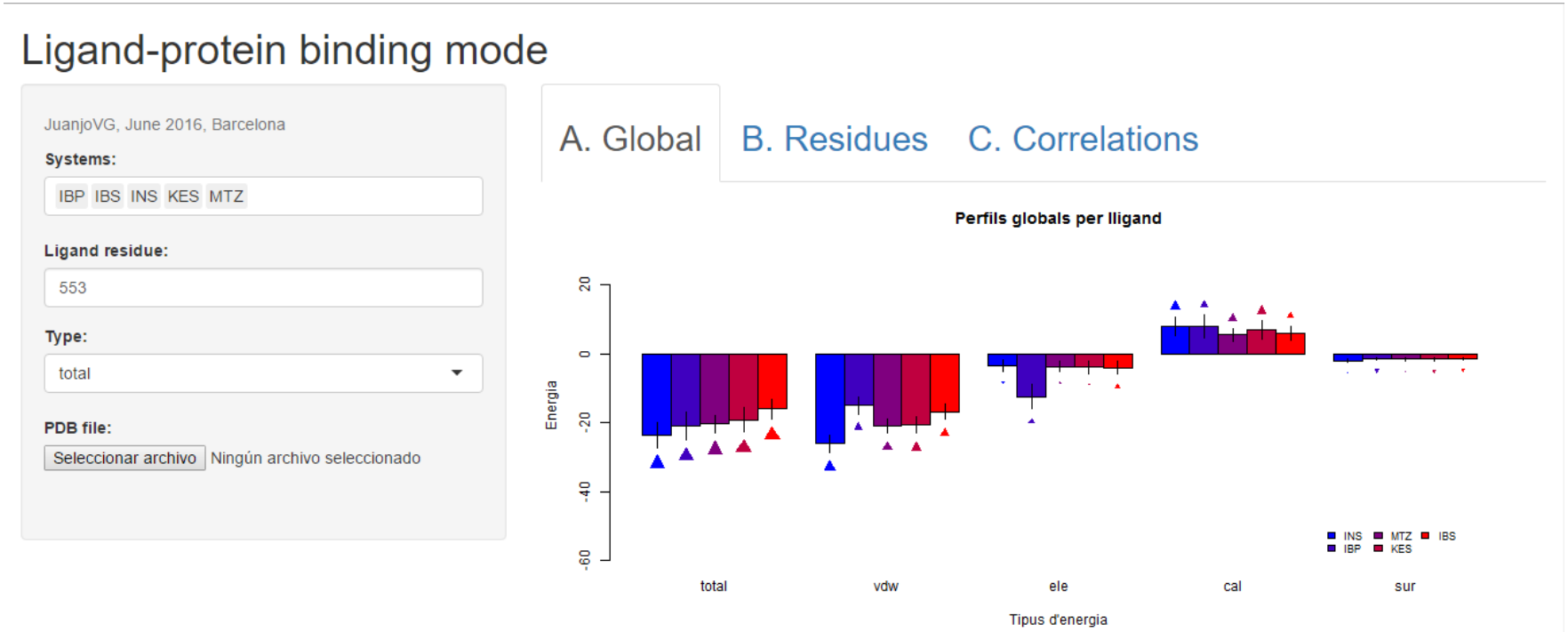


Figura 10.13: Aspecte de l'aplicació web a la vista de perfils energetics per residu.

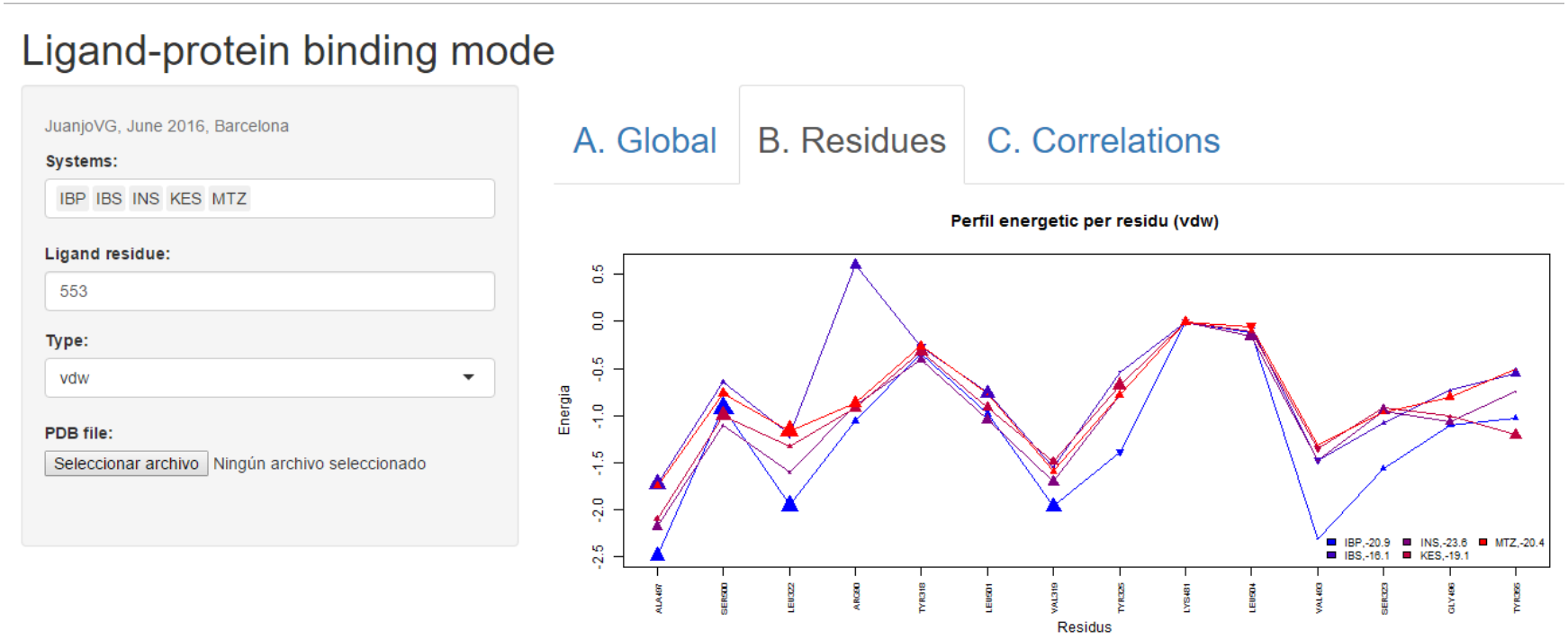


Figura 10.14: Aspecte de l'aplicació web a la vista de correlacions sense la informació dels àtoms més propers.

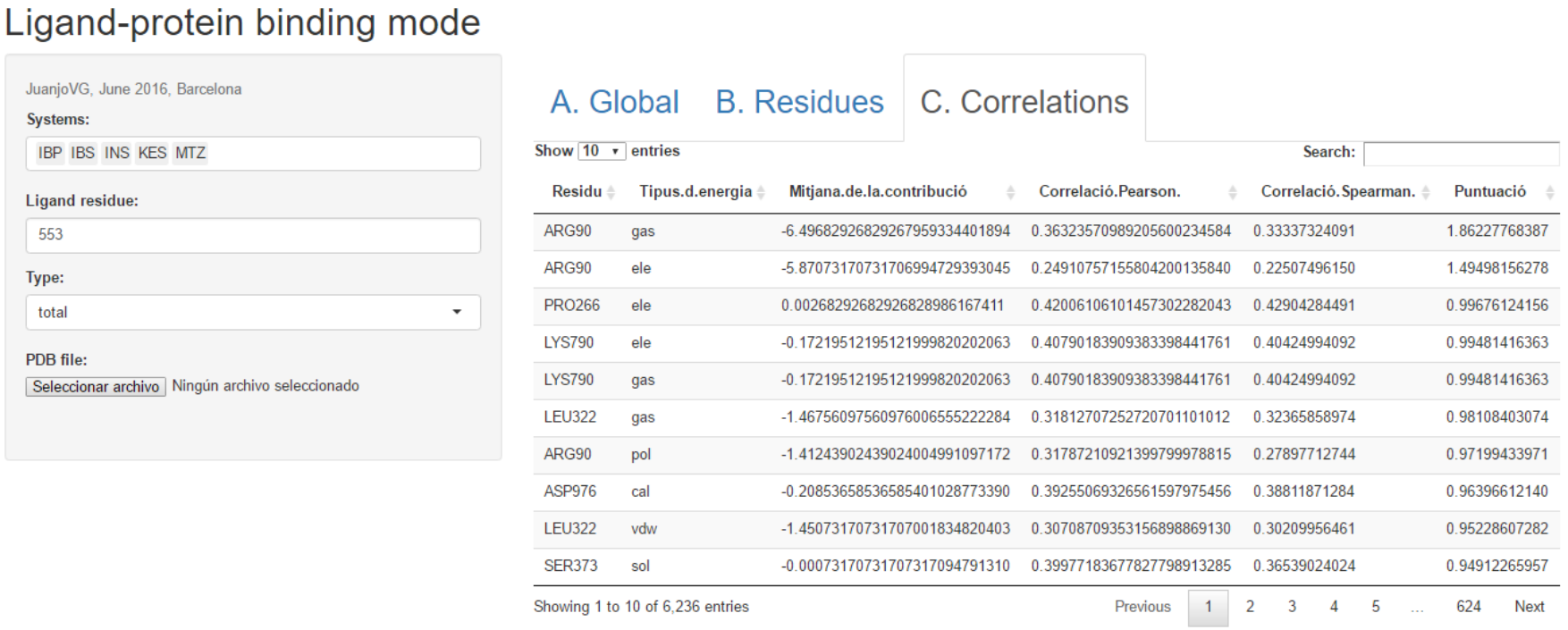


Figura 10.15: Aspecte de l'aplicació web a la vista de correlacions amb la informació dels àtoms més propers.